Niveau(x) d'enseignement: 0 Numéro EDUQ : 8670353
Centre d'intérêt: 090

Allaire, Denis.

Théorie statistique des scores de distance : études préliminaires / Denis Allaire. -- [Sainte-Foy] : IHRS-Education, c1975.
64, [10] f. : graph. ; 28 cm.

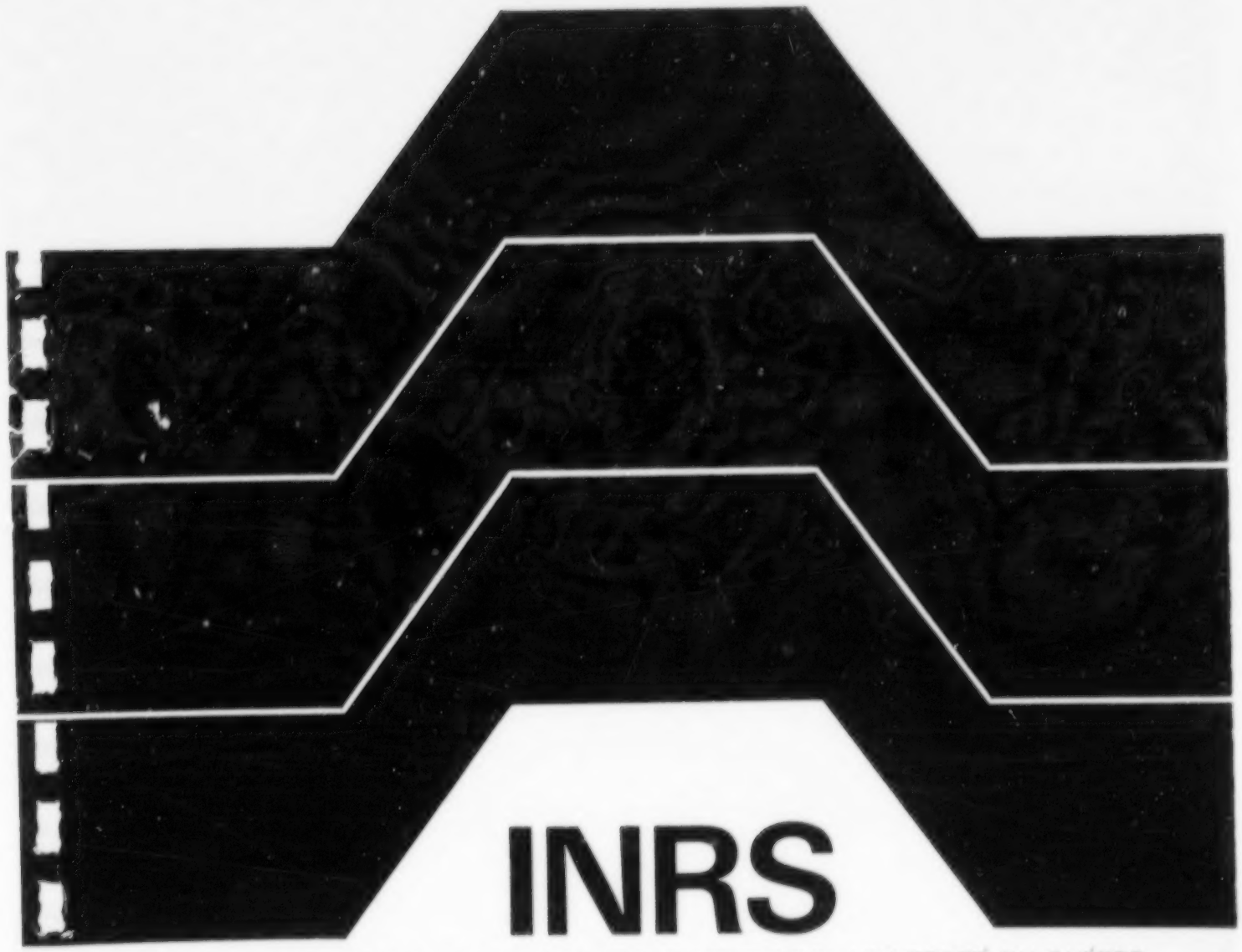
Code: Document D-12. Bibliogr.: f. [74].

Vedette(s)-secondaire(s): IHRS-Education

Descripteur(s): Analyse de variables * Analyse statistique * Validité * Statistique

De façon générale, un score de distance permet de mesurer à quel point deux individus (ou deux objets) sont relativement semblables ou différents sur une ou plusieurs variables. Le document contient sept courtes études visant à élucider certains aspects des scores de distance et, principalement, leur fidélité: 1. comparaison entre trois méthodes de calculer la distance. Etude de simulation; 2. moyenne, variance et fidélité des scores de distance; 3. postulats impliqués par la formule de la fidélité des distances; 4. statistiques usuelles dans le cas de variables normales élevées au carré; 5. la fidélité des scores de distance dans le cas de variables reliées. Etude de simulation; 6. la fidélité des distances calculées sur plusieurs variables normales quelconques; 7. la fidélité et la validité des distances calculées après sommation de plusieurs variables quelconques.

Type de document: r




INSTITUT NATIONAL DE LA RECHERCHE SCIENTIFIQUE - UNIVERSITÉ DU QUÉBEC

EDUQ 8670353

# Théorie statistique des scores de distance: études préliminaires . I

Volume I

Denis Allaire

INRS-Education

12 mars 1975

© Denis Allaire, 1975.

Document D-12

Nous avons réuni en ces pages quelques courtes études visant à élucider certains aspects des scores de distance et, principalement, leur fidélité.

De façon générale, un score de distance permet de mesurer à quel point deux individus (ou deux objets) sont relativement semblables ou différents sur une ou plusieurs variables. La distance entre les individus j et j' sur k variables d'intérêt se symbolise par $d_{jj'(k)}$ et s'obtient par l'équation 1.

$$d_{jj'(k)} = \sqrt{\sum_{i=1}^{k} (x_{ij} - x_{ij'})^2} \qquad (1)$$

Il ne s'agit en fait que d'une généralisation du théorème de Pythagore. La distance ainsi calculée s'interprète, en termes géométriques, comme la distance entre les positions des individus j et j' dans un espace à k dimensions orthogonales, positions dont les k coordonnées correspondent aux k scores obtenus par chaque individu. Dans certains cas, on pourra préférer utiliser le carré de $d_{jj'}$, soit $d^2_{jj'}$, à cause de ses avantages mathématiques, lesquels sont comparables aux avantages de la variance par rapport à l'écart-type.

La mesure de distance n'est pas la seule disponible afin de mesurer le degré de ressemblance entre individus. Un grand nombre de mesures de "similarité entre profils" a été proposé. Toutes ces mesures appartiennent de façon générale à deux grandes familles, selon qu'elles s'inspirent soit de la formule de la distance, soit de la formule de la corrélation. Cronbach (1953) et Cronbach et Gleser (1953) ont déjà produit une excellente discussion de bon nombre de ces formules. Certaines études ont également été réalisées afin d'en comparer les mérites respectifs (Helmstadter, 1957; Mosel et Roberts, 1954; Carroll et Field, 1974). Afin de mieux approfondir l'une d'elles, soit la mesure de distance, nous avons cru bon d'ignorer les autres mesures dans une première étape.

## Table des matières


| | | page |
|---|---|---|
| Etude 1 - | Comparaison entre trois méthodes de calculer la distance. Etude de simulation | 3 |
| Etude 2 - | Moyenne, variance et fidélité des scores de distance | 11 |
| Etude 3 - | Postulats impliqués par la formule de la fidélité des distances | 17 |
| Etude 4 - | Statistiques usuelles dans le cas de variables normales élevées au carré | 21 |
| Etude 5 - | La fidélité des scores de distances dans le cas de variables reliées. Etude de simulation | 28 |
| Etude 6 - | La fidélité des distances calculées sur plusieurs variables normales quelconques | 47 |
| Etude 7 - | La fidélité et la validité des distances calculées après sommation de plusieurs variables quelconques | 53 |

## ETUDE 1

### Comparaison entre trois méthodes de calculer la distance Etude de simulation

Supposons qu'un chercheur ait recueilli p mesures sur n individus et qu'il soit intéressé, pour un motif quelconque, à mesurer la distance existant entre tous ces individus pris deux à deux. Une question qui risque de se présenter à lui immédiatement est celle de savoir s'il considérera, dans le calcul des $d_{jj'}$, les p variables originales ou si, au contraire, il n'en utilisera pas un nombre moindre en opérant le regroupement de certaines. Par exemple, ces p mesures pourraient être les réponses à chacun des items d'un test relativement homogène. Est-il préférable de réunir ces items en un même score ou de les traiter séparément dans le calcul des $d_{jj'}$? Est-ce que ces alternatives s'avèrent comparables ou, au contraire, l'une ne serait-elle pas supérieure en termes de validité ou de fidélité? Autre exemple, les p mesures sont constituées de sous-groupes censés mesurer r domaines relativement distincts. Ici, trois méthodes de calculer la distance s'avèrent possibles: a) séparément sur les p scores originaux; b) sur les r scores redevables au regroupement des mesures partageant un même domaine, et c) sur un seul score issu du regroupement de l'ensemble des mesures.

Cette question du regroupement possible de certaines mesures a déjà été soulevée par Cronbach et Gleser (1953, cf. pp. 470-471). Leur discussion peut se résumer ainsi. Quand on regroupe des variables, cela a toujours pour effet d'accorder un plus grand poids à ce qu'elles ont en commun par rapport à ce qu'elles ont de spécifique. C'est l'effet contraire qui se produit si on les considère séparément: la contribution de leur partie spécifique s'en trouve augmentée. Il s'en suit que si des variables sont supposées, pour des raisons logiques ou statistiques, mesurer un même trait, il serait préférable de les regrouper. Dans ce cas, la part spécifique de ces mesures est assimilable à de l'erreur de mesure, laquelle verra son importance diminuer suite au regroupement. Si, au contraire, les variables sont supposées mesurer des domaines distincts, il devient préférable de les considérer séparément afin que leurs variances spécifiques

conservent un certain poids dans l'attribution des distances. Procéder autrement entraînerait ici une contribution exagérée du "facteur général" par rapport aux "facteurs spécifiques".

A l'appui de leurs énoncés, Cronbach et Gleser citent des résultats obtenus par Warrington (1952) au moyen de données simulant les réponses aux items d'un (ou de plusieurs?) test homogène. Le regroupement des items dans un seul score aurait produit ici une augmentation considérable de la validité des scores de distance. Ces résultats apparaissent d'autant plus précieux qu'ils confirment des conclusions qui n'étaient, par ailleurs, étayées par aucun résultat, ni par aucun développement formel. Malheureusement, cette recherche de Warrington est difficilement accessible et le résumé qu'en font Cronbach et Gleser est insuffisamment explicite pour qu'on puisse se faire une idée juste de la portée des résultats obtenus.

Afin de pallier à cette absence de résultats détaillés, nous avons réalisé une étude de Monte Carlo visant à vérifier les conclusions de Cronbach et Gleser, principalement:

a) que la réunion de variables mesurant le même domaine augmente la fidélité des scores de distance;

b) que la réunion de variables mesurant des domaines distincts diminue la fidélité des scores de distance.

## Description de l'expérience

Afin de vérifier ces hypothèses, il est nécessaire de créer un certain nombre de variables faillibles mesurant des domaines distincts. Pour ce faire, les scores vrais des sujets sur deux facteurs ont d'abord été créés à partir d'une population normale. A chacun, cinq composantes d'erreur ont ensuite été ajoutées afin de produire cinq scores parallèles d'un niveau donné de fidélité. Les distances vraies entre sujets ont été obtenues à partir des scores factoriels vrais de chaque sujet. Trois méthodes ont été utilisées afin de calculer les distances observées:

1) sur les dix scores considérés séparément;

2) sur deux scores dus au regroupement des cinq premières variables et des cinq dernières;

3) sur un seul score dû au regroupement des dix variables.

La première de nos hypothèses veut que les distances calculées par la seconde méthode (k = 2) soient en plus forte corrélation avec les distances vraies que celles calculées par la première méthode (k = 10) puisqu'il y a eu réunion de variables homogènes. La seconde hypothèse veut que les distances de la troisième méthode (k = 1) soient moins fidèles que celles de la seconde (k = 2) puisque des variables indépendantes ont été réunies.

La fidélité[1] de ces trois méthodes a été observée sur 20 échantillons de 50 sujets. Pour chaque échantillon, l'expérience a été reprise trois fois en variant le niveau de fidélité des dix variables servant au calcul des distances. Des fidélités de .06, .20 et .50 ont été produites en variant le poids accordé aux mêmes dix composantes d'erreur. L'appendice A produit le programme ayant servi à la génération de ces résultats.

## Résultats

Le Tableau 1 présente les corrélations entre distances vraies et distances observées, obtenues par chacun des 20 échantillons pour chacune des neuf conditions expérimentales (3 méthodes x 3 fidélités) ainsi que leurs moyennes. Ces corrélations ont été calculées sur les 1225 paires différentes de sujets de chaque échantillon (n (n - 1)/2). Les distances n'ont pas été élevées au carré afin que leur distribution ne soit pas trop asymétrique.

Ces corrélations, après avoir été transformées en z de Fisher, ont été soumises à une analyse de variance dont les résultats apparaissent au Tableau 2.

---

[1] Plus précisément, il s'agit de la racine carrée de la fidélité puisque la corrélation entre scores observés et scores vrais n'a pas été élevée au carré.

Tableau 1

Corrélations entre distances vraies et distances observées dans les neuf conditions expérimentales

| Echantillons (n = 50) | nombre de dimensions considérées | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | k = 10 | | | k = 2 | | | k = 1 | | |
| | r = .06 | .20 | .50 | r = .06 | .20 | .50 | r = .06 | .20 | .50 |
| 1 | .15 | .39 | .74 | .24 | .52 | .80 | .08 | .30 | .53 |
| 2 | .24 | .50 | .81 | .36 | .63 | .86 | .24 | .46 | .63 |
| 3 | .15 | .46 | .81 | .30 | .60 | .86 | .16 | .34 | .52 |
| 4 | .21 | .46 | .79 | .19 | .56 | .84 | .05 | .25 | .44 |
| 5 | .02 | .36 | .77 | .23 | .55 | .83 | .12 | .35 | .54 |
| 6 | .04 | .29 | .71 | .16 | .49 | .80 | .17 | .39 | .57 |
| 7 | .13 | .43 | .77 | .33 | .59 | .83 | .25 | .41 | .55 |
| 8 | .18 | .46 | .81 | .33 | .61 | .85 | .11 | .34 | .55 |
| 9 | .07 | .36 | .77 | .17 | .50 | .82 | .08 | .29 | .54 |
| 10 | .10 | .38 | .76 | .23 | .53 | .82 | .06 | .28 | .51 |
| 11 | .26 | .52 | .81 | .32 | .62 | .85 | .15 | .32 | .49 |
| 12 | .28 | .52 | .80 | .26 | .56 | .82 | .17 | .44 | .65 |
| 13 | .23 | .50 | .80 | .38 | .63 | .85 | .31 | .50 | .63 |
| 14 | -.06 | .17 | .65 | .02 | .36 | .76 | .03 | .28 | .55 |
| 15 | .17 | .46 | .81 | .37 | .64 | .86 | .21 | .37 | .51 |
| 16 | .06 | .35 | .74 | .21 | .53 | .81 | .13 | .33 | .54 |
| 17 | .10 | .39 | .77 | .14 | .49 | .80 | .09 | .33 | .56 |
| 18 | .11 | .37 | .74 | .13 | .46 | .79 | .10 | .35 | .58 |
| 19 | -.02 | .21 | .68 | .06 | .40 | .77 | .05 | .29 | .54 |
| 20 | .17 | .44 | .77 | .26 | .54 | .81 | .17 | .36 | .55 |
| moyenne[a] | .130 | .406 | .769 | .236 | .545 | .824 | .137 | .351 | .550 |

[a] Ces moyennes ont été calculées au moyen des transformations z de Fisher.

La Figure 1 illustre les effets les plus intéressants de cette analyse. Pour ce qui est de l'effet associé aux trois méthodes de calcul des distances ($F_{2,38} = 144.5$; $p < .01$), on y voit clairement que la méthode réunissant les variables en deux groupes (k = 2) s'avère la plus fidèle. La supériorité de cette méthode sur les deux autres (k = 10 et k = 1) vérifie donc les deux hypothèses que nous avions formulées: regrouper les variables appartenant à un même facteur augmente la fidélité, mais regrouper celles appartenant à des facteurs distincts la diminue.

Quant à l'effet de la fidélité des variables sur la fidélité des distances ($F_{2,38} = 7569.7$; $p < .01$), il apparaît également très nettement

Tableau 2

Analyse de variances des corrélations
obtenues après transformation
en z de Fisher

| Source de variance | S.C. | d.l. | C.M. | F |
|---|---|---|---|---|
| Méth. | 24,882.311 | 2 | 13,441.16 | 144.54** |
| Fid. | 178,946.711 | 2 | 89,473.36 | 7,569.66** |
| Ech. | 10,370.550 | 19 | 545.82 | 58.63[a]** |
| M x F | 13,455.022 | 4 | 3,363.76 | 361.31** |
| M x E | 3,533.461 | 38 | 92.99 | 9.99[a]** |
| F x E | 449.066 | 38 | 11.82 | 1.27[a] |
| Res | 707.866 | 76 | 9.31 | |
| Totale | 234,344.994 | 179 | 1,309.19 | |

** significatif à .01

[a] Ces rapports F sont des estimés inférieurs puisque la triple interaction a dû être utilisée comme terme d'erreur.

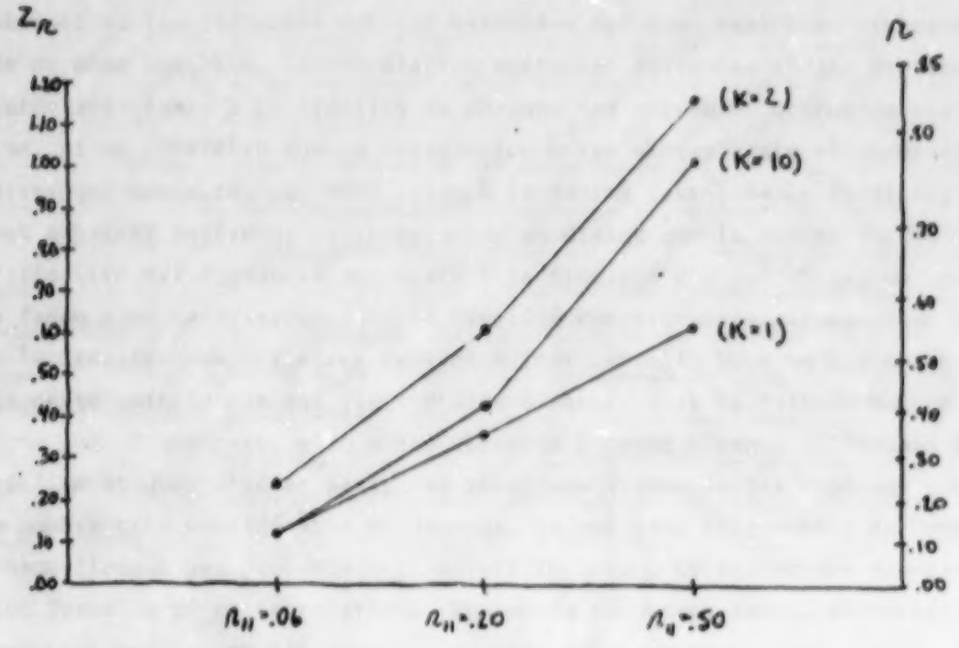



Figure 1 - Corrélations moyennes entre distances vraies et distances observées pour trois mesures de distance et trois niveaux de fidélité des variables.

à la Figure 1. Il est tout-à-fait normal que la fidélité des distances aille croissant quand la fidélité des mesures dont elles proviennent augmente. Il est difficile de préciser la forme exacte de la fonction qui unit ces deux fidélités, mais on sait au moins quels en sont les deux points extrêmes. Si la fidélité des variables est nulle, celle des distances le sera également. Si la fidélité des variables est parfaite, celle des distances le sera aussi. A propos de ce lien entre fidélité des variables et fidélité des distances, une coincidence intéressante, présente dans ces résultats, doit être notée. Dans le cas des distances dues à la réunion des variables en deux groupes (k = 2), on remarque que les corrélations moyennes avec les distances vraies (.236, .545 et .824) correspondent presqu'exactement à la fidélité de chacun des deux regroupements.

Ces regroupements étant la somme de cinq mesures parallèles de même fidélité, leur fidélité, prédite au moyen de la formule de Spearman-Brown, sera de .238, .556 et .833, tout dépendant du niveau initial de fidélité (.06, .20 et .50). Il semblerait donc que, dans un cas comme celui-ci où les distances ont été calculées sur deux variables orthogonales de même fidélité, la corrélation entre ces distances et les distances vraies soit égale à la fidélité de chacune des variables orthogonales. De plus, si on considère que la corrélation entre scores vrais et observés correspond toujours, par définition, à la racine carrée de la fidélité, on peut exprimer autrement cette relation en disant que la racine carrée de la fidélité des distances est égale à la fidélité des variables; ou encore, de façon plus saisissante, que la fidélité des distances est égale au carré de la fidélité des variables servant à leur calcul! Nous ne croyons pas que cette coincidence qui vient d'être discutée soit le fait du hasard parce que, d'une part, elle a été observée à trois niveaux différents de fidélité et que, d'autre part, les résultats présentés ici reposent sur un nombre très considérable de données, ce qui rend l'hypothèse du hasard échantillonnal peu probable. Il serait important de tenter une démonstration formelle de cette relation. Également important serait de vérifier, analytiquement ou empiriquement, si cette relation vaut aussi dans les cas où le nombre de variables orthogonales est différent de deux.

Un troisième effet qui ressort significatif de l'analyse de variance du Tableau 1 est celui redevable aux différences entre échantillons ($F_{19,76} = 58.6$; $p < .01$). Cet effet apparaît tout-à-fait normal. Les 20 échantillons au hasard qui ont été utilisés étant de grandeur moyenne ($n = 50$), il est inévitable que la variance des scores vrais et celles des scores d'erreur varient quelque peu de l'un à l'autre. Ces différences de variances ont nécessairement une influence sur les coefficients de corrélation servant de variable dépendante dans cette étude. Puisque chacun des échantillons a servi dans les neuf conditions expérimentales, il aurait été étonnant que les différences existantes entre eux n'atteignent pas le seuil de la signification.

Un autre effet significatif est celui de l'interaction entre le niveau de fidélité et la méthode servant au calcul des distances ($F_{4,76} = 361.3$; $p < .01$). Cette interaction, comme l'illustre la Figure 1, provient du non-parallélisme des effets simples, Ce non-parallélisme est d'ailleurs une caractéristique obligatoire des courbes de la Figure 1, puisque toutes trois doivent nécessairement converger à l'origine des axes dans le cas limite où la fidélité des variables serait nulle.

Un dernier effet significatif révélé par l'analyse de variance met en cause l'interaction entre les méthodes et les échantillons ($F_{38,76} = 9.9$; $p < .01$). Il est difficile d'expliquer pourquoi la façon de calculer les distances devrait avoir un effet différent d'un échantillon à l'autre. La seule hypothèse qui nous est apparue défendable ici est que la corrélation entre les deux scores vrais ait suffisamment varié autour de sa valeur attendue de zéro pour que, dans le cas de certains échantillons, la supériorité relative de la distance utilisant deux regroupements ($k = 2$) s'en trouve modifiée. Cette hypothèse ne peut cependant pas être vérifiée ici puisque les corrélations échantillonnales entre scores vrais n'ont pas été enregistrées. De toutes façons, cette interaction s'avère l'effet le moins important parmi ceux qui ont été discutés.

## Discussion et conclusion

Parmi les résultats précédents, le plus important est, sans contredit, la supériorité qui a été observée de la part d'une mesure de distance qui est calculée en réunissant les variables parallèles et en séparant les variables distinctes. Les présentes données démontrent le bien-fondé de la discussion que Cronbach et Gleser avaient faite de ce problème et la justesse de leurs conclusions.

Les résultats obtenus ici proviennent évidemment d'une situation de mesure idéale. Toutes les variables étaient de même fidélité et elles étaient soit rigoureusement parallèles, soit rigoureusement orthogonales. Les situations réelles ne sont pas en général aussi tranchées. Les variables ne sont pas parfaitement parallèles, ni factoriellement pures. Les variables observées seront au mieux constituées de sous-groupes relativement distincts à l'intérieur desquels les variables ne seront que relativement homogènes. Comme prolongement à cette étude, il serait intéressant de voir ce qu'il en est des trois méthodes expérimentées ici dans le cas de facteurs qui ne sont pas orthogonaux mais qui présentent des degrés divers de relation. Sur la foi des présents résultats, on peut d'ores et déjà prévoir que, dans le cas extrême, où les facteurs seraient en corrélation parfaite, la méthode regroupant toutes les variables en une seule ($k = 1$) surpasserait celle qui ne réunissait que les variables d'un même domaine ($k = 2$). Reste à savoir ce qu'il en serait de ces deux méthodes dans les autres cas intermédiaires.

Un autre résultat important de cette étude, pour lequel aucune hypothèse n'avait initialement été formulée, est que la fidélité des distances est égale au carré de la fidélité des variables, du moins dans le cas où les distances observées ont été calculées sur deux variables indépendantes de même fidélité. Cette relation entre les deux niveaux de fidélité, celle des variables et celle des distances, mériterait d'être étudiée plus avant.

## ETUDE 2

### Moyenne, variance et fidélité des scores de distance

Nous disions en conclusion à l'étude précédente que "cette relation entre les deux niveaux de fidélité, celle des variables et celle des distances, mériterait d'être étudiée plus avant". Dans les résultats obtenus par simulation, la fidélité des distances était en effet apparue égale au carré de la fidélité des variables d'origine. C'est dans l'espoir de démontrer cette relation que la présente étude a été entreprise.

Afin de réussir à exprimer la fidélité des distances en fonction de celle des variables, il était d'abord nécessaire d'exprimer la variance des distances en fonction de celle des variables. A son tour, la formule de la variance nécessitait que la moyenne des distances puisse aussi être exprimée en fonction des variables d'origine. Ces formules, celles de la moyenne, de la variance et de la fidélité des scores de distance, nous avons réussi à les produire, mais uniquement dans deux des trois situations de mesure qui avaient été envisagées. La première situation est celle d'une seule variable. La seconde est celle où les distances sont calculées sur plusieurs variables indépendantes. La troisième situation, celle où nous avons rencontré des difficultés, correspond au cas plus général - et aussi plus réel - de plusieurs variables dépendantes.

Il est à noter que toutes les formules que nous présentons ne valent que pour les scores de distance élevés au carré, soit les $d^2_{jj'}$, et non pour les distances simples. Nous sommes loin d'être persuadé que les mêmes formules puissent être produites dans le cas des $d_{jj'}$. Cette limitation apparaît malheureuse si on considère que les distances simples peuvent s'avérer préférables aux distances au carré dans maintes utilisations de ces scores, particulièrement dans les cas où la plus grande normalité des $d_{jj'}$ est requise. A propos de cette disparité entre $d_{jj'}$ et $d^2_{jj'}$, il se pourrait bien que ces mesures, finalement, ne soient pas aussi différentes qu'on serait porté à le croire. Nous mentionnerons, à

titre d'exemple, cette relation entre fidélité des distances et celle des variables, dont nous démontrons ici la formule dans le cas des $d^2_{jj'}$, mais qui avait pourtant été observée sur des $d_{jj'}$ dans les résultats de simulation précédents.

## 1. Le cas d'une seule dimension

Soit X, une variable qui se distribue selon $N(\mu_X, \sigma^2_X)$ dans la population.

Soit une nouvelle variable D, obtenue en échantillonnant au hasard deux valeurs de la population des X et en faisant la différence

$$D = (X - X') \tag{2}$$

Les différences D auront donc une distribution normale de moyenne zéro et de variance $2\sigma^2_X$, i.e. D sera $N(0, 2\sigma^2_X)$.

Si la variable D est divisée par son écart-type, cette variable transformée $Z_D$

$$Z_D = D/\sqrt{2\sigma^2_X} \tag{3}$$

sera N(0,1). Ceci permet d'exprimer D comme une transformation linéaire d'une variable normale centrée-réduite:

$$D = \sqrt{2\sigma^2_X} \cdot Z_D \tag{4}$$

Cette dernière expression permet d'obtenir aisément les moments de la distribution des D en fonction des moments connus de la courbe normale qui sont:

$$\begin{aligned} E(Z) &= 0, \quad E(Z^2) = 1 \\ E(Z^3) &= 0, \quad E(Z^4) = 3 \end{aligned} \tag{5}$$

Les moments de D seront donc:

$$\begin{aligned} E(D) &= E\left(\sqrt{2\sigma^2_X} \cdot Z_D\right) \\ &= \sqrt{2\sigma^2_X} \cdot E(Z) = 0 \end{aligned} \tag{6}$$

$$\begin{aligned} E(D^2) &= E\left[2\sigma^2_X \cdot Z^2_D\right] \\ &= 2\sigma^2_X \cdot E(Z^2) = 2\sigma^2_X \end{aligned} \tag{7}$$

$$E(D^3) = E\left[(2\sigma_X^2)^{3/2} \cdot Z_D^3\right]$$
$$= (2\sigma_X^2)^{3/2} \cdot E(Z^3) = 0 \qquad (8)$$

$$E(D^4) = E\left[(2\sigma_X^2)^2 \cdot Z_D^4\right]$$
$$= 4\sigma_X^4 \cdot E(Z^4) = 12(\sigma_X^2)^2 \qquad (9)$$

Les moments de D étant liés à la variance de X, il est possible d'exprimer la moyenne et la variance des $D^2$ en fonction de la variance des X. Ainsi

$$\mu(D^2) = E(D^2) = 2\sigma_X^2 \qquad (10)$$

et

$$\sigma^2(D^2) = E\left[(D^2)^2\right] - \left[E(D^2)\right]^2$$
$$= E(D^4) - \left[E(D^2)\right]^2$$
$$= 12(\sigma_X^2)^2 - (2\sigma_X^2)^2 = 8(\sigma_X^2)^2 \qquad (11)$$

Les différences au carré, $D^2$, ont donc une moyenne égale au double de la variance des X, et une variance égale à huit fois la variance des X élevée au carré.

A partir de la variance des $D^2$, il est désormais possible de trouver une expression à leur fidélité, pourvu que soit adoptée la définition classique du "rapport de la variance vraie sur la variance observée". Soit X les scores vrais et x les scores faillibles auxquels correspondent les différences vraies D et les différences faillibles d, alors

$$\rho^2(d^2, D^2) = \sigma^2(D^2) / \sigma^2(d^2)$$
$$= 8(\sigma_X^2)^2 / 8(\sigma_x^2)^2$$
$$= (\sigma_X^2 / \sigma_x^2)^2 = \left[\rho^2(x,X)\right]^2 \qquad (12)$$

Autrement dit, la fidélité des $d^2$ est égale au carré de la fidélité des x.

## 2. Le cas de r dimensions orthogonales

Soit r variables orthogonales normales, $X_1$, $X_2$, ...$X_i$...$X_r$, de

moyennes $\mu_1, \mu_2, \ldots \mu_i \ldots \mu_r$ et de variances $\sigma_1^2, \sigma_2^2 \ldots \sigma_i^2 \ldots \sigma_r^2$.

Soit la variable des distances, $D^2$, obtenue en échantillonnant au hasard deux vecteurs de r valeurs et en en faisant la somme des différences au carré

$$D^2_{(r)} = \sum_{i=1}^{r} (X_i - X'_i)^2 = \sum_{i=1}^{r} D_i^2 \qquad (13)$$

La moyenne des $D^2_{(r)}$ sera égale à la somme des moyennes des $D^2$ propres à chaque variable, soit[1]

$$\mu(D^2_{(r)}) = \sum_{i=1}^{r} \left[\mu(D_i^2)\right]$$

$$= 2 \sum_{i=1}^{r} \sigma^2(X_i) \qquad (14)$$

Autrement dit, la moyenne des distances, $D^2_{(r)}$, dans un espace à r dimensions orthogonales, est égale à deux fois la somme des variances de ces dimensions.

La variance des $D^2_{(r)}$ est égale à la somme des variances des $D^2$ propres à chaque dimension, puisque les dimensions sont indépendantes et que leurs covariances sont conséquemment nulles.

$$\sigma^2(D^2_{(r)}) = \sum_{i=1}^{r} \left[\sigma^2(D_i^2)\right]$$

$$= \sum_{i=1}^{r} \left[ 8(\sigma^2_{X_i})^2 \right]$$

$$= 8 \sum_{i=1}^{r} (\sigma^2 X_i)^2 \qquad (15)$$

Autrement dit, la variance des distances, dans un espace à r dimensions orthogonales, est égale à huit fois la somme du carré des variances associées à chacune des dimensions.

Quant à la fidélité des distances observées, $d^2_{(r)}$, elle s'obtient par

$$\rho^2(d^2_{(r)}, D^2_{(r)}) = \sigma^2(D^2_{(r)}) \,/\, \sigma^2(d^2_{(r)})$$

---

[1] Cette formule avait déjà été présentée par Cronbach et Gleser (1953, p. 471).

$$= 8 \sum_{i=1}^{r} \sigma^4(X_i) \; / \; 8 \sum^{r} \sigma^4(x_i)$$

$$= \sum^{r} \sigma^4(X_i) \; / \; \sum^{r} \sigma^4(x_i) \qquad (16)$$

Si toutes les variables ont la même variance et la même fidélité, l'expression précédente se simplifie et on obtient

$$\rho^2(d^2_{(r)}, D^2_{(r)}) = \sigma^4(X_i) \; / \; \sigma^4(x_i) = \left[\rho^2(x_i, X_i)\right]^2 \qquad (17)$$

Autrement dit, dans le cas de r dimensions orthogonales comme dans le cas d'une seule dimension, la fidélité des distances est égale au carré de la fidélité des variables, pourvu que ces dernières aient des variances et des fidélités identiques.

### 3. Le cas de r dimensions obliques

Ce cas est celui dont l'intérêt pratique est le plus grand puisque les variables réelles utilisées dans le calcul des distances seront rarement orthogonales les unes par rapport aux autres.

La moyenne de $D^2_{(r)}$ est la même que dans le cas précédent et s'obtient par l'équation (14).

Contrairement au cas précédent, la variance de $D^2_{(r)}$ n'est plus ici la somme des variances associées à chaque $D^2_i$. Ces r différences $D^2_i$ doivent nécessairement être en corréiation puisque les r variables dont elles proviennent le sont. La variance de $D^2_{(r)}$ sera donc égale à la somme des éléments de la matrice variance-covariance entre les r $D^2_i$, soit

$$\sigma^2(D^2_{(r)}) = \sum_{i=1}^{r} \sum_{i'=1}^{r} \sigma^2(D^2_i, D^2_{i'}) \qquad (18)$$

Malheureusement, contrairement aux deux cas précédents, nous avons vainement tenté d'exprimer cette variance en fonction des variables d'origine Nous croyons cependant qu'une simplification en est possible. La solution nécessite que la covariance entre les différences $D^2_i$ et $D^2_j$, provenant de deux variables $X_i$ et $X_j$, puisse être exprimée en fonction des variances de $X_i$ et $X_j$ et de leur covariance. En espérant qu'un lecteur éventuel

puisse obtenir la formule recherchée, nous présentons ci-dessous ce qui pourrait être le début d'une solution.

$$\sigma(D_i^2,\ D_j^2) = \sigma\left[2\sigma^2(X_i) \cdot Z_i^2,\ 2\sigma^2(X_j) \cdot Z_j^2\right]$$

$$= 4\sigma^2(X_i)\ \sigma^2(X_j)\ \sigma(Z_i^2,\ Z_j^2)$$

$$\text{où } \sigma(Z_i^2,\ Z_j^2) = E\ (Z_i^2\ Z_j^2) - E\ (Z_i^2)\ E\ (Z_j^2)$$

$$= E\ (Z_i^2\ Z_j^2) - 1$$

Si $E\ (Z_i^2\ Z_j^2)$ pouvait être exprimé en fonction de $E\ (Z_i\ Z_j)$, comme cela nous apparaît possible intuitivement, il n'y aurait plus alors qu'un pas, facile à franchir, jusqu'à la solution recherchée.

En l'absence d'une formule de la variance des distances dans le cas oblique, il devient également impossible de trouver une expression à leur fidélité.

## ETUDE 3

### Postulats impliqués par la formule de la fidélité des distances

Dans l'étude précédente, nous avons présenté une expression à la fidélité des distances, $d^2$, dans le cas d'une seule dimension ou de plusieurs dimensions orthogonales. Ces formules (12 et 16), nous les avons obtenues en utilisant une des définitions classiques de la fidélité, soit celle du rapport entre variances vraie et observée.

De façon à nous conformer à l'usage courant, nous avons également symbolisé ces fidélités par $\rho^2(d^2, D^2)$, soit le carré de la corrélation entre distances observées et distances vraies. Ce faisant, nous établissions une égalité sans pour autant la démontrer. Le but de cette étude est de vérifier si cette égalité entre la fidélité "rapport-de-variance" et la fidélité "corrélation-au-carré" vaut pour les scores de distance comme elle vaut pour les scores construits dans la théorie des tests.

\- - -

Soit les distances observées, d, définies comme les différences entre les scores observés x et x' échantillonnés au hasard à partir de la même distribution.

$$d = x - x' \qquad (1)$$

Les scores observés x et x' sont définis, conformément à la théorie des tests, comme la somme d'un score vrai et d'un score d'erreur, soit

$$x = X + E \qquad (2a)$$
$$x' = X' + E' \qquad (2b)$$

Alors, les distances observées d peuvent également être définies comme la somme d'une "distance vraie" et d'une "distance d'erreur", soit

$$d = X + E - X' - E' = D + Ed \qquad (3)$$

où $D = X - X'$ et $Ed = E - E'$

Pour démontrer que "le rapport de la variance des distances vraies sur la variance des distances observées" est égale au "carré de la corrélation entre distances vraies et observées", il est nécessaire que la "covariance entre distances vraies et observées" soit égale à la "variance des distances vraies". En effet, si

$$\sigma(d^2, D^2) = \sigma^2(D^2) \qquad (4)$$

alors
$$\rho^2(d^2, D^2) = \sigma^2(d^2, D^2) \,/\, \sigma^2(d^2)\sigma^2(D^2)$$
$$= \sigma^4(D^2) \,/\, \sigma^2(d^2)\sigma^2(D^2)$$
$$= \sigma^2(D^2) \,/\, \sigma^2(d^2) \qquad (5)$$

Mais, afin de vérifier l'équation 4, il est nécessaire de postuler ici l'indépendance expérimentale[1] entre les composantes vraie et d'erreur, D et Ed, constituant les scores d. Dans le cas de la théorie classique des tests, le seul postulat nécessaire était l'indépendance linéaire entre composantes vraie et d'erreur du score observé, soit X et E. En effet

$$\sigma(x,X) = \sigma(X + E, X) = \sigma^2(X) + \sigma(XE)$$
$$= \sigma^2(X) + \mathcal{E}XE - \mathcal{E}X\mathcal{E}E = \sigma^2(X)$$

parce que l'indépendance linéaire implique que $\mathcal{E}XE = \mathcal{E}X\mathcal{E}E$. Ce postulat n'est cependant pas suffisant, dans le cas des scores de distance, pour résoudre l'équation 4. En effet,

$$\sigma(d^2, D^2) = \sigma\left[(D + Ed)^2, D^2\right] = \sigma(D^2 + E^2d + 2DEd, D^2)$$
$$= \sigma^2(D^2) + \sigma(D^2, E_d^2) + 2\sigma(D^2, DEd) \qquad (6)$$

Cependant, si D et Ed sont expérimentalement indépendants, les deux derniers termes de l'expression de droite disparaîtront. Ainsi

$$\sigma(D^2, E_d^2) = \mathcal{E}D^2E_d^2 - \mathcal{E}D^2\mathcal{E}E_d^2 = 0 \qquad (7)$$

et
$$\sigma(D^2, DEd) = \mathcal{E}D^3Ed - \mathcal{E}D^2\mathcal{E}DEd$$
$$= \mathcal{E}D^3\mathcal{E}Ed - \mathcal{E}D^2\mathcal{E}D\mathcal{E}Ed$$
$$= \mathcal{E}Ed(\mathcal{E}D^3 - \mathcal{E}D^2\mathcal{E}D)$$

---

[1] L'indépendance linéaire a comme seule implication que $\mathcal{E}AB = \mathcal{E}A\mathcal{E}B$. Par contre l'indépendance expérimentale se définit par le fait que la distribution conjointe de A et B est égale au produit de leur distribution respective. Cela implique, entre autres choses, que

$\mathcal{E}A^uB^v = \mathcal{E}A^u\mathcal{E}B^v$ pour tout u et v.

$$= \mathcal{E}_{Ed}\left[\sigma(D^2, D)\right] = 0 \qquad (8)$$

parce que $\mathcal{E}_{Ed} = 0$.

A son tour, l'indépendance expérimentale entre D et $E_D$ nécessite que, non seulement x et x' soient expérimentalement indépendants, mais aussi X et E. Pour nous en convaincre, supposons que x et x' sont expérimentalement indépendants, mais que X et E, d'une part, et X' et E', d'autre part, ne sont que linéairement indépendants. Il s'en suit que X et X', E et E', X et E' ainsi que X' et E seront aussi expérimentalement indépendants. Cependant l'ensemble de ces conditions n'est pas suffisant pour démontrer l'indépendance expérimentale de D et Ed, ni même pour démontrer leur <u>indépendance quadratique</u>[1], puisque

$$\mathcal{E}_{D^2Ed^2} - \mathcal{E}_{D^2}\mathcal{E}_{Ed^2} = \mathcal{E}(X-X')^2(E-E')^2 - \mathcal{E}(X-X')^2\,\mathcal{E}(E-E')^2$$

$$= \mathcal{E}\left[(X^2+X'^2-2XX')\,(E^2+E'^2-2EE')\right]$$

$$- (\mathcal{E}X^2+\mathcal{E}X'^2-2\mathcal{E}XX')\,(\mathcal{E}E^2+\mathcal{E}E'^2-2\mathcal{E}EE')$$

$$= a+b+c+d+e+f+g+h+i = a+b+g+h$$

où (a) $= \mathcal{E}X^2E^2 - \mathcal{E}X^2\mathcal{E}E^2$

(b) $= \mathcal{E}X'^2E'^2 - \mathcal{E}X'^2\mathcal{E}E'^2$

(c) $= \mathcal{E}X^2E'^2 - \mathcal{E}X^2\mathcal{E}E'^2 = 0$

(d) $= \mathcal{E}X'^2E^2 - \mathcal{E}X'^2\mathcal{E}E^2 = 0$

(e) $= -2(\mathcal{E}X^2EE' - \mathcal{E}X^2\mathcal{E}EE') = -2(\mathcal{E}X^2E\mathcal{E}E' - \mathcal{E}X^2\mathcal{E}E\mathcal{E}E') = 0$

(f) $= -2(\mathcal{E}X'^2EE' - \mathcal{E}X'^2\mathcal{E}EE') = -2(\mathcal{E}X'^2E'\mathcal{E}E - \mathcal{E}X'^2\mathcal{E}E\mathcal{E}E') = 0$

(g) $= -2(\mathcal{E}E^2XX' - \mathcal{E}E^2\mathcal{E}XX') = -2(\mathcal{E}E^2X\mathcal{E}X' - \mathcal{E}E^2\mathcal{E}X\mathcal{E}X')$

(h) $= -2(\mathcal{E}E'^2XX' - \mathcal{E}E'^2\mathcal{E}XX') = -2(\mathcal{E}E'^2X'\mathcal{E}X - \mathcal{E}E'^2\mathcal{E}X\,X')$

(i) $= +4(\mathcal{E}XX'EE' - \mathcal{E}XX'\mathcal{E}EE') = +4(\mathcal{E}XE\mathcal{E}X'E' - \mathcal{E}X\mathcal{E}X'\mathcal{E}E\mathcal{E}E')$

$$= +4(\mathcal{E}X\mathcal{E}E\mathcal{E}X'\mathcal{E}E' - \mathcal{E}X\mathcal{E}E\mathcal{E}X'\mathcal{E}E') = 0$$

---

[1] Comme l'indépendance linéaire, l'indépendance quadratique est un cas particulier de l'indépendance expérimentale. Elle se définit par

$$\mathcal{E}A^uB^v = \mathcal{E}A^u\mathcal{E}B^v \qquad \text{pour u et v} \leq 2$$

Le seul postulat de l'indépendance linéaire permet d'annuler l'équation (i). L'indépendance expérimentale entre x et x' permet d'annuler (c) et d). Ce dernier postulat ajouté à celui de $\mathcal{C}E = 0$ permet d'annuler (e) et (f). Par contre, les équations (a) et (b) ne peuvent être annulées qu'à la condition que

$$\mathcal{C}X^2E^2 = \mathcal{C}X^2\mathcal{C}E^2,$$

et les équations (g) et (h), qu'à la condition que

$$\mathcal{C}XE^2 = \mathcal{C}X\mathcal{C}E^2$$

Ces deux dernières conditions nécessitant l'indépendance quadratique de X et E, il s'en suit que le postulat classique de l'indépendance linéaire de X et E ne constitue pas une condition suffisante à la démonstration de l'équation 7, sans laquelle on ne peut vérifier l'équivalence, dans le cas des scores de distance, entre les deux définitions classiques de la fidélité (cf. équation 5).

De la même façon, l'équation 8 pourrait aussi être développée afin de montrer que l'indépendance cubique entre X et E constitue également une condition nécessaire à la vérification de l'équation 5.

Les développements précédents permettent donc d'avancer la conclusion suivante. Dans le cas des scores de distances, comme dans le cas des scores usuels de la théorie classique des tests, le carré de la corrélation entre scores vrais et scores observés est égal au rapport de la variance vraie sur la variance observée. Cependant, contrairement au cas des scores classiques, ce résultat nécessite que les distances proviennent de scores dont les composantes vraie et d'erreur sont expérimentalement indépendantes, ou, de façon minimum, cubiquement indépendantes. Ce résultat nécessite aussi que les scores x et x', dont on fait la différence, soient eux-mêmes expérimentalement indépendants. Dans le cas où x et x' sont échantillonnés au hasard, ce dernier postulat est évidemment respecté.

En aucun temps, les développements précédents n'ont nécessité le postulat de normalité. Ce postulat s'avère cependant indispensable si on veut exprimer la fidélité des distances en fonction de la fidélité des scores dont elles proviennent, comme nous l'avons fait dans l'étude 2.

## ETUDE 4

### Statistiques usuelles dans le cas de variables normales élevées au carré

L'étude 2 présentait un certain nombre de formules se rapportant aux scores de distance. Les scores de distance n'étant eux-mêmes, du moins dans le cas d'une seule variable, que des scores de différences élevées au carré, il apparaît donc que les formules de l'étude 2 correspondent à un cas particulier de scores élevés au carré, soit celui des scores de différence.

Nous avons cru utile, dans cette nouvelle étude, de développer les formules de la moyenne, de la variance, de la covariance et de la corrélation pour le cas, plus général, de variables normales quelconques (section A). Les formules correspondant aux cas plus simples de variables centrées ou centrées réduites sont également présentées (sections B et C). Le cas particulier des scores de distance est finalement traité comme un cas spécial de variables centrées (section D).

Il est bon de se rappeler que les formules que nous présentons, exception faite de celle de la moyenne, n'auraient pu être simplifiées dans leurs formes actuelles sans que la normalité des variables ne soit postulée. L'avantage de ce postulat est de permettre que les moments d'ordre 3 et 4 d'une distribution soient exprimés en fonction de ses deux premiers moments. Dans le cas de variables dont la forme de la distribution n'est pas connue, les formules correspondantes nécessiteraient que les moments d'ordre 3 et 4 soient directement obtenus. Leur expression en serait donc beaucoup plus complexe.

Sous la forme du Tableau 1, nous présentons un résumé des formules les plus utiles de cette étude.

#### A. Le cas général

Soit une variable normale X exprimée comme une transformation linéaire d'une variable normale centrée réduite Z

$$X = \sigma Z + \mu \qquad (1)$$

Tableau 1

Formulaire

| Statistiques | X / Y | | $U_x = X-\mu_x$ / $U_y = Y-\mu_y$ | | $Z_x = (X-\mu_x)/\sigma_x$ / $Z_y = (Y-\mu_y)/\sigma_y$ | | $D_x = X - X'$ / $D_y = Y - Y'$ | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{E}X$ | $= \mu_x$ | (2) | 0 | (22) | 0 | (4) | 0 | (34) |
| $\mathcal{E}X^2$ | $= \sigma^2_x - \mu^2_x$ | (3) | $\sigma^2_x$ | (23) | 1 | (5) | $2\sigma^2_x$ | (37) |
| $\mathcal{E}X^3$ | $= 3\sigma^2_x \mu_x + \mu^3_x$ | (8) | 0 | (24) | 0 | (6) | 0 | (38) |
| $\mathcal{E}X^4$ | $= 3\sigma^4_x + 6\sigma^2_x \mu^2_x + \mu^4_x$ | (9) | $3\sigma^4_x$ | (25) | 3 | (7) | $12\sigma^4_x$ | (39) |
| $\sigma^2(X)$ | $= \sigma^2_x$ | | $\sigma^2_x$ | | 1 | | $2\sigma^2_x$ | (35) |
| $\sigma^2(X^2)$ | $= 2\sigma^4_x + 4\sigma^2_x \mu^2_x$ | (10) | $2\sigma^4_x$ | (26) | 2 | (30) | $8\sigma^4_x$ | (40) |
| $\sigma(X,Y)$ | $= \sigma_{xy}$ | | $\sigma_{xy}$ | | $\rho_{xy}$ | | $2\sigma_{xy}$ | (36) |
| $\sigma(X^2,Y^2)$ | $= 2\sigma^2_{xy} + 4\sigma_{xy} \mu_x \mu_y$ | (19) | $2\sigma^2_{xy}$ | (27) | $2\rho^2_{xy}$ | (31) | $8\sigma^2_{xy}$ | (41) |
| $\rho(X^2,Y^2)$ | $= \dfrac{\sigma^2_{xy} + 2\sigma_{xy} \mu_x \mu_y}{(\sigma^4_x \sigma^4_y + 2\sigma^4_x \sigma^2_y \mu^2_y + 2\sigma^2_x \sigma^4_y \mu^2_x + 4\sigma^2_x \sigma^2_y \mu_x \mu_y)^{1/2}}$ | (20) | $\rho^2_{xy}$ | (28) | $\rho^2_{xy}$ | (32) | $\rho^2_{xy}$ | (42) |

Les deux premiers moments de X sont

$$\mathcal{E}X = \mu \tag{2}$$

$$\mathcal{E}X^2 = \sigma^2 + \mu^2 \tag{3}$$

Connaissant les moments de la distribution normale centrée réduite,

$$\mathcal{E}z = 0 \tag{4}$$

$$\mathcal{E}z^2 = 1 \tag{5}$$

$$\mathcal{E}z^3 = 0 \tag{6}$$

$$\mathcal{E}z^4 = 3 \tag{7}$$

il est aisé d'obtenir les autres moments de X

$$\begin{aligned}\mathcal{E}X^3 &= \mathcal{E}(\sigma z + \mu)^3 \\ &= \mathcal{E}(\sigma^3 z + 3\mu\sigma^2 z^2 + 3\mu^2\sigma z + \mu^3) \\ &= \sigma^3\mathcal{E}z^3 + 3\mu\sigma^2\mathcal{E}z^2 + 3\mu^2\sigma\mathcal{E}z + \mu^3 \\ &= 0 + 3\mu\sigma^2 + 0 + \mu^3 \\ &= 3\mu\sigma^2 + \mu^3 \end{aligned} \tag{8}$$

$$\begin{aligned}\mathcal{E}X^4 &= \mathcal{E}(\sigma z + \mu)^4 \\ &= \mathcal{E}(\sigma^4 z^4 + 4\mu\sigma^3 z^3 + 6\mu^2\sigma^2 z^2 + 4\mu^3\sigma z + \mu^4) \\ &= \sigma^4\mathcal{E}z^4 + 4\mu\sigma^3\mathcal{E}z^3 + 6\mu^2\sigma^2\mathcal{E}z^2 + 4\mu^3\sigma\mathcal{E}z + \mu^4 \\ &= 3\sigma^4 + 0 + 6\mu^2\sigma^2 + 0 + \mu^4 \\ &= 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4 \end{aligned} \tag{9}$$

Il s'en suit que la variance des valeurs au carré s'obtient par

$$\begin{aligned}\sigma^2(X^2) &= \mathcal{E}(X^2)^2 - \left[\mathcal{E}(X^2)\right]^2 \\ &= \mathcal{E}X^4 - (\mathcal{E}X^2)^2 \\ &= 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4 - (\sigma^2 + \mu^2)^2 \\ &= 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4 - \sigma^4 - \mu^4 - 2\sigma^2\mu^2 \\ &= 2\sigma^4 + 4\mu^2\sigma^2 \end{aligned} \tag{10}$$

Soit les variables normales X et Y. Afin d'exprimer la covariance entre $X^2$ et $Y^2$, il est nécessaire de décomposer X et Y en une composante commune, C, et en une composante expérimentalement indépendante de la première, soit I et J respectivement

$$X = C + I \tag{11a}$$

$$Y = C + J \tag{11b}$$

Les composantes I et J étant indépendantes d'une même troisième, elles sont donc, par transitivité indépendantes entre elles. Les composantes C, I et J étant mutuellement expérimentalement indépendantes, il s'en suit, par définition, que

$$\mathcal{E}C^qI^rJ^s = \mathcal{E}C^q\mathcal{E}I^r\mathcal{E}J^s \tag{12}$$

Alors

$$\sigma(X,Y) = \sigma(C + I, C + J) = \sigma^2(C) \tag{13}$$

parce que, par l'équation 12, $\sigma(C,I) = \sigma(C,J) = \sigma(I,J) = 0$.

Quant à la covariance entre les carrés de X et Y, elle s'obtient par

$$\begin{aligned}\sigma(X^2,Y^2) &= \sigma\left[(C^2 + I^2 + 2CI),(C^2 + J^2 + 2CJ)\right] \\ &= \sigma^2(C^2) + \sigma(C^2,I^2) + \sigma(C^2,J^2) + \sigma(I^2,J^2) + 2\sigma(I^2,CJ) \\ &\quad + 2\sigma(J^2,CI) + 2\sigma(C^2,CI) + 2\sigma(C^2,CJ) + 4\sigma(CI,CJ) \\ &= \sigma^2(C^2) + 2\sigma(C^2,CI) + 2\sigma(C^2,CJ) + 4\sigma(CI,CJ)\end{aligned} \tag{14}$$

parce que, par l'équation 12, les cinq autres termes s'annulent. Mais, par l'équation 10,

$$\sigma^2(C^2) = 2\sigma^2(C)\left[\sigma^2(C) + 2(\mathcal{E}C)^2\right] \tag{15}$$

tandis que, par les équations 3 et 8,

$$\begin{aligned}\sigma(C^2,CI) &= \mathcal{E}C^3I - \mathcal{E}C^2\mathcal{E}CI \\ &= \mathcal{E}C^3\mathcal{E}I - \mathcal{E}C^2\mathcal{E}C\mathcal{E}I \\ &= \mathcal{E}I(\mathcal{E}C^3 - \mathcal{E}C^2\mathcal{E}C) \\ &= \mathcal{E}I\left[3\sigma^2(C)\mathcal{E}C - (\mathcal{E}C)^3 - \sigma^2(C)\mathcal{E}C + (\mathcal{E}C)^3\right] \\ &= 2\sigma^2(C)\mathcal{E}C\mathcal{E}I\end{aligned} \tag{16}$$

et, de la même façon,

$$\sigma(C^2,CJ) = 2\sigma^2(C)\ell_C\ell_J \qquad (17)$$

et, finalement,

$$\begin{aligned}\sigma(CI,CJ) &= \ell C^2_{IJ} - \ell_C\ell_I\ell_C\ell_J \\ &= \ell C^2\ell_I\ell_J - (\ell C)^2\ell_I\ell_J \\ &= \sigma^2(C)\ell_I\ell_J \qquad (18)\end{aligned}$$

En utilisant les équations 15 à 18, la covariance de l'équation 14 se simplifie en

$$\begin{aligned}\sigma(X^2,Y^2) &= 2\sigma^2(C)\left[\sigma^2(C) + 2(\ell C)^2 + 2\ell_C\ell_I + 2\ell_C\ell_J + 2\ell_I\ell_J\right] \\ &= 2\sigma^2(C)\left[\sigma^2(C) + 2(\ell_C + \ell_I)(\ell_C + \ell_J)\right] \\ &= 2\sigma^2(C)\left[\sigma^2(C) + 2\ell_X\ell_Y\right] \\ &= 2\sigma^2(X,Y) + 4\sigma(X,Y)\ell_X\ell_Y \qquad (19)\end{aligned}$$

puisque $\sigma(X,Y) = \sigma^2(C)$, par l'équation 13, et que, suite aux définitions 11a et 11b, $X = \ell_C + \ell_I$ et $\ell_Y = \ell_C + \ell_J$.

Quant à la corrélation entre $X^2$ et $Y^2$, elle s'obtient au moyen des équations 10 et 19,

$$\begin{aligned}\rho(X^2,Y^2) &= \sigma(X^2,Y^2) \,/\, \left[\sigma^2(X^2)\sigma^2(Y^2)\right]^{1/2} \\ &= \sigma(X^2,Y^2) \,/\, \left[(2\sigma^4_x + 4\sigma^2_x\mu^2_x)(2\sigma^4_y + 4\sigma^2_y\mu^2_y)\right]^{1/2} \\ &= \frac{\sigma^2(X,Y) + 2\sigma(X,Y)\mu_x\mu_y}{(\sigma^4_x\sigma^4_y + 2\sigma^4_x\sigma^2_y\mu^2_y + 2\sigma^2_x\sigma^4_y\mu^2_x + 4\sigma^2_x\sigma^2_y\mu^2_x\mu^2_y)^{1/2}} \qquad (20)\end{aligned}$$

B. Le cas de variables centrées

Soit les variables normales centrées $U_x$ et $U_y$

$$U_x = X - \mu_x \qquad (21a)$$

$$U_y = Y - \mu_y \qquad (21b)$$

Puisque, dans ce cas, $\mathcal{E}(U_x) = \mathcal{E}(U_y) = 0$, il est aisé de simplifier les formules précédentes. Ainsi

$$\mathcal{E}U_x = 0 \tag{22}$$

$$\mathcal{E}U_x^2 = \sigma^2_x \tag{23}$$

$$\mathcal{E}U_x^3 = 0 \tag{24}$$

$$\mathcal{E}U_x^4 = 3\sigma^4_x \tag{25}$$

$$\sigma^2(U_x^2) = 2\sigma^4_x \tag{26}$$

$$\sigma(U_x^2, U_y^2) = 2\sigma^2_{xy} \tag{27}$$

$$\rho(U_x^2, U_y^2) = \sigma^2_{xy}/\sigma^4_x\sigma^4_y = \rho^2_{xy} \tag{28}$$

C. Le cas de variables centrées réduites

Soit les variables centrées réduites $Z_x$ et $Z_y$

$$Z_x = (X - \mu_x)/\sigma_x \tag{29a}$$

$$Z_y = (Y - \mu_y)/\sigma_y \tag{29b}$$

Puisque, dans ce cas particulier, les moyennes sont nulles et les variances égales à l'unité, il est aisé de simplifier encore les formules précédentes.

$$\sigma^2(Z_x^2) = 2 \tag{30}$$

$$\sigma(Z_x^2, Z_y^2) = 2\rho^2_{xy} \tag{31}$$

$$(Z_x^2, Z_y^2) = \rho^2_{xy} \tag{32}$$

D. Le cas des distances

Soit les valeurs X et X' échantillonnées au hasard de la distribution normale de X, et les valeurs Y et Y' échantillonnées au hasard de la distribution normale de Y. Soit les distances ou différences $D_x$ et $D_y$ définies comme

$$D_x = X - X' \tag{33a}$$

$$D_y = Y - Y' \tag{33b}$$

Les variables X et X' étant échantillonnées d'une même distribution, il s'en suit que

$$\mathcal{E}D_x = \mathcal{E}X - \mathcal{E}X' = 0 \qquad (34)$$

Les variables X et X' étant échantillonnées au hasard, elles sont donc expérimentalement indépendantes; il s'en suit que leur covariance est nulle. Alors

$$\sigma^2(D_x) = \sigma^2_x + \sigma^2_{x'} - 2\sigma(X,X') \qquad (35)$$
$$= 2\sigma^2_x$$

et

$$\sigma(D_x,D_y) = \sigma(X - X', Y - Y')$$
$$= \sigma(X,Y) - \sigma(X,Y') - \sigma(X',Y) + \sigma(X',Y')$$
$$= 2\sigma_{xy} \qquad (36)$$

Il est à noter que les variables $D_x$ et $D_y$ sont elles-mêmes normales, puisqu'elles proviennent de différences entre variances normales.

Les formules correspondant aux distances au carré, $D^2$, s'obtiennent aisément en appliquant les équations 35 et 36 aux formules traitant du cas des variables centrées, soit les équations 23 à 28. Ainsi,

$$\mathcal{E}D^2_x = \sigma^2(D_x) = 2\sigma^2_x \qquad (37)$$

$$\mathcal{E}D^3_x = 0 \qquad (38)$$

$$\mathcal{E}D^4_x = 3\left[\sigma^2(D_x)\right]^2 = 12\sigma^4_x \qquad (39)$$

$$\sigma^2(D_x) = 2\left[\sigma^2(D_x)\right]^2 = 8\sigma^4_x \qquad (40)$$

$$\sigma(D^2_x,D^2_y) = 2\left[2\sigma(D_x,D_y)\right]^2 = 8\sigma^2_{xy} \qquad (41)$$

$$\rho(D^2_x,D^2_y) = \sigma(D^2_x,D^2_y)\Big/\left[\sigma^2(D^2_x)\sigma^2(D^2_y)\right]^{1/2}$$
$$= 8\sigma^2_{xy} \Big/ (8\sigma^4_x \cdot 8\sigma^4_y)^{1/2}$$
$$= \sigma^2_{xy} \Big/ \sigma^2_x\sigma^2_y$$
$$= \rho^2_{xy} \qquad (42)$$

## ETUDE 5

### La fidélité des scores de distances dans le cas de variables reliées Etude de simulation

Les résultats de simulation de l'étude 1 avaient clairement montré que les mesures de distances étaient plus fidèles quand, préalablement à leur calcul, on regroupait les variables parallèles (scores vrais identiques) et qu'on séparait les variables distinctes (scores vrais indépendants). A titre d'exemple, imaginons que, pour chaque individu, nous ayons deux scores observés, chacun étant la somme d'un score vrai et d'un score d'erreur aléatoire. Si les scores vrais sont indépendants (corrélation nulle), mieux vaut calculer les distances en considérant séparément les deux scores observés. La corrélation entre les distances calculées sur les deux scores vrais et celles calculées sur les deux scores observés sera alors maximum ($K = 2$). Par contre, si les deux scores vrais sont les mêmes (corrélation parfaite), mieux vaut faire la somme des deux scores observés. La corrélation entre les distances calculées sur les deux scores vrais et celles calculées sur un seul score observé - dû à la sommation des deux scores initiaux - sera maximum ($K = 1$).

Le but de cette nouvelle étude est d'explorer, comme l'idée en avait été suggérée en conclusion à l'étude 1, ce qu'il advient de la fidélité de ces deux méthodes ($K = 1$ et $K = 2$) dans les cas intermédiaires où la corrélation entre scores vrais n'est ni nulle, ni parfaite. Plus précisément, le but est ici de comparer les fonctions unissant la fidélité des distances à la corrélation entre scores vrais, dans le cas de l'une et l'autre méthode.

Les résultats de l'étude 1 ainsi que certaines équations de l'étude 2 permettent d'ores et déjà de prévoir certains aspects de ces deux fonctions. Une première prédiction est que ces fonctions devront se croiser à un quelconque niveau de corrélation entre scores vrais. Ce fait, illustré à la Figure 1 découle de la supériorité de l'une des méthodes ($K = 2$) dans le cas extrême où les scores vrais sont indépendants

et de la supériorité de l'autre méthode (K = 1) dans le cas opposé où les scores vrais sont identiques. Une autre prédiction, également illustrée à la Figure 1, concerne le point de départ de l'une de ces fonctions et le point d'arrivée de l'autre.

Dans le cas d'une corrélation nulle, les distances calculées sur deux variables auront une fidélité égale au carré de la fidélité de ces variables (étude 2, équation 17). A l'autre extrémité, quand les scores vrais seront en corrélation parfaite et que les deux dimensions ne feront plus qu'une, les distances calculées sur une seule variable - due à la sommation des deux variables initiales - auront aussi une fidélité égale au carré de la fidélité de cette variable (étude 2, équation 12 ). Puisque, dans ce dernier cas, les scores vrais sont identiques et, par voie de conséquence, les variables sont parallèles, la fidélité de cette somme peut être prédite par la formule de Spearman-Brown.

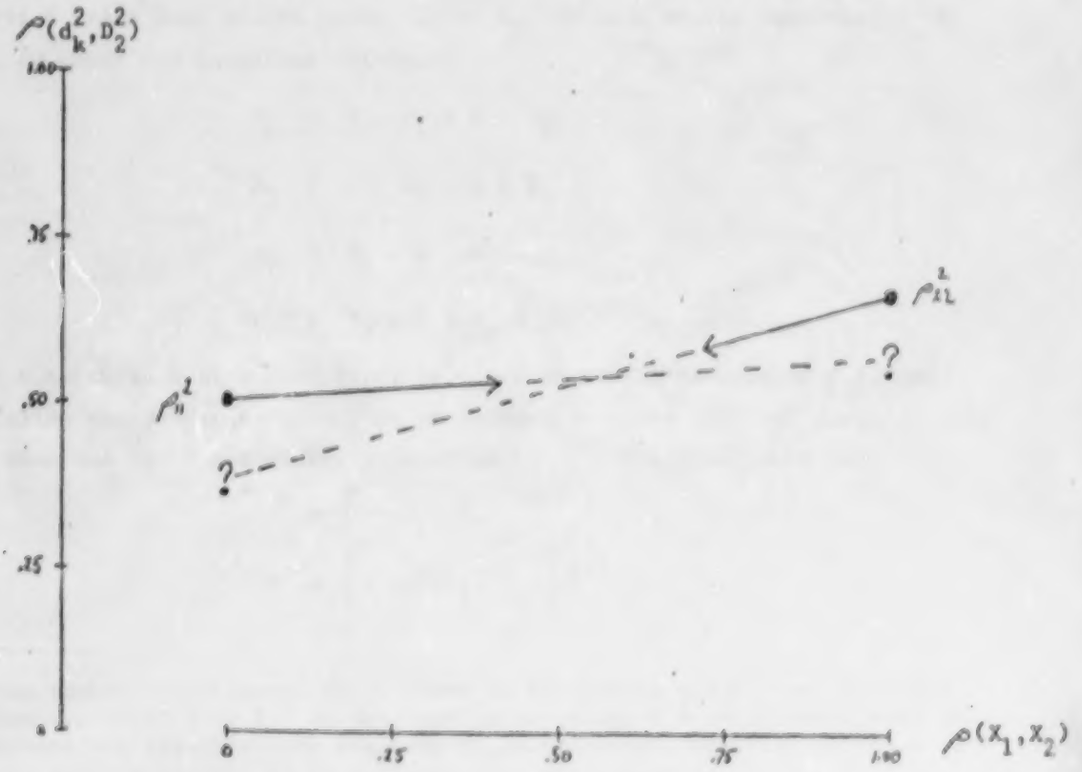



Figure 1. Illustration de certaines prédictions concernant les fonctions unissant la fidélité des distances à la corrélation entre scores vrais.

Ces faits étant établis, il reste encore à savoir quelles formes précises auront ces fonctions et à quel niveau elles se croiseront. La réponse à cette dernière question comporte un intérêt pratique évident. Les variables réelles étant rarement, soit parfaitement parallèles, soit parfaitement indépendantes, il est utile de connaître à partir de quel niveau de corrélation il devient avantageux, dans le calcul des distances, de les réunir ou de les séparer.

## Expérience 1

### 1. Méthode de simulation

Les données de cette expérience ont été obtenues de la façon suivante. Dans un premier temps, cinq nombres aléatoires normaux[1], $Z_1$, $Z_2$, ...$Z_5$, ont été générés pour chaque sujet (n = 500). Ces nombres ont ensuite servi à créer deux scores vrais, $X_1$ et $X_2$, et deux scores observés, $x_1$ et $x_2$, au moyen des équations suivantes

$$X_1 = A \cdot Z_1 + B \cdot Z_2$$

$$X_2 = A \cdot Z_1 + B \cdot Z_3$$

$$x_1 = X_1 + C \cdot Z_4$$

$$x_2 = X_2 + C \cdot Z_5$$

Les constantes A et B permettent de créer un niveau prédéterminé de corrélation entre scores vrais tout en faisant en sorte que leur variance soit la même que celle des scores aléatoires Z. Elles s'obtiennent par

$$A = \left[\rho(X_1, X_2)\right]^{1/2}$$

$$B = \left[1 - \rho(X_1, X_2)\right]^{1/2}$$

---

[1] Ces nombres proviennent de la somme de 12 nombres aléatoires uniformes compris entre 0 et 1. En soustrayant la valeur 6 à cette somme, nous obtenons une variable très légèrement platykurtique, comprise entre -6 et +6, dont la moyenne et la variance sont exactement égales à 0 et à 1.

Quant à la constante C, elle permet de créer un niveau prédéterminé de fidélité. Dans un cas comme celui-ci où scores vrais et composantes d'erreur (i.e. $Z_4$ ou $Z_5$) ont la même variance, elle s'obtient par

$$C = \left[1/\rho^2 (x_1, X_1) - 1\right]^{1/2}$$

## 2. Schème expérimental

La fidélité des distances ou, plus précisément, la corrélation entre les distances calculées sur les scores vrais et celles calculées sur les scores observés (variable dépendante) a été observée pour l'une et l'autre méthode (K = 1 et K = 2), à cinq niveaux de corrélation entre scores vrais ($\rho(X_1,X_2)$ = .00, .25, .50, .75 et 1.00) et à deux niveaux de fidélité des variables ($\rho^2(x_1, X_1)$ = .06 et .50), pour un grand total de 20 conditions expérimentales. Pour chacune de ces conditions, dix échantillons de 50 sujets ont été utilisés. Les cinq nombres aléatoires servant à créer les scores vrais et observés de chaque sujet demeuraient les mêmes dans toutes les conditions expérimentales et seules les trois constantes A, B et C voyaient leur valeur modifiée. Cette procédure a été adoptée afin de minimiser l'influence des fluctuations échantillonnales d'une condition expérimentale à l'autre.

## 3. Résultats

Pour chaque condition expérimentale, la corrélation entre distances vraies et observées a été obtenue pour chacun des dix échantillons. Chacune de ces corrélations était calculée sur les 1225 paires différentes de sujets qu'un échantillon de 50 sujets permet d'obtenir. Les corrélations utilisaient non pas les distances au carré, mais leur racine carrée (i.e. $\sqrt{D^2}$), afin que leur distribution s'approche un peu plus de la normale.

La Figure 2 présente l'essentiel des résultats de cette expérience sous la forme de quatre fonctions en traits continus. Chaque point qui y est représenté est la corrélation moyenne de dix échantillons. Ces corrélations moyennes, présentées au Tableau 1, ont été obtenues par l'intermédiaire des transformations z de Fisher.

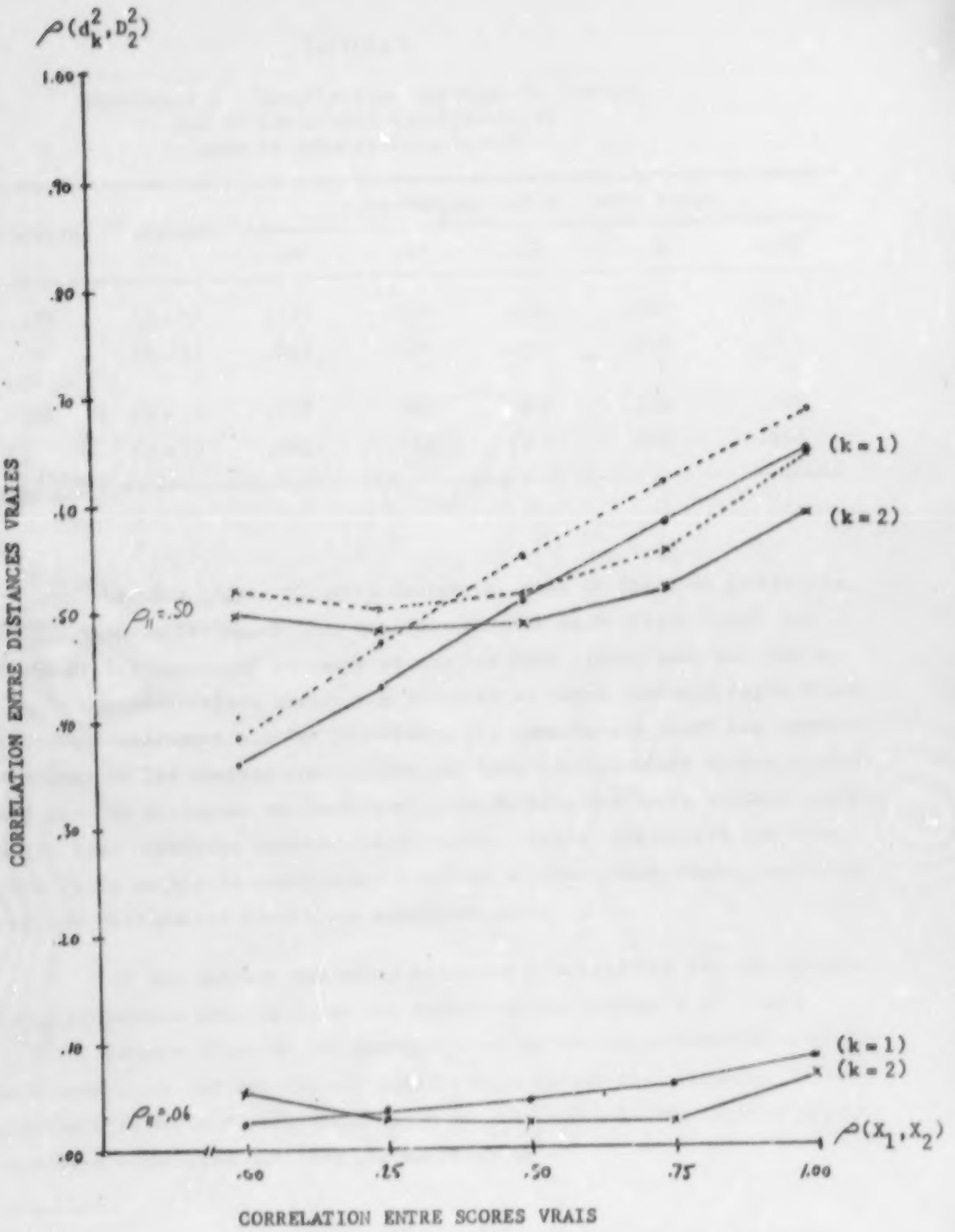



Figure 2. Résultats de l'expérience 1

Tableau 1

Expérience 1 - Corrélations moyennes de chacune des 20 conditions expérimentales pour 10 échantillons (n = 50)

| Fidélité | Méthode | corrélations entre scores vrais | | | | |
|---|---|---|---|---|---|---|
| | | .00 | .25 | .50 | .75 | 1.00 |
| .06 | (K = 2) | .054 | .029 | .026 | .034 | .067 |
| " | (K = 1) | .025 | .034 | .044 | .058 | .083 |
| .50 | (K = 2) | .499 | .482 | .489 | .520 | .589 |
| " | (K = 1) | .360 | .432 | .510 | .583 | .648 |

La même figure illustre également, sous la forme de pointillés, les fonctions obtenues, à l'un des deux niveaux de fidélité (.50), en utilisant les distances au carré et non pas leur racine carrée. Ces résultats supplémentaires visaient à vérifier si cette variante était d'une quelconque influence sur les résultats. La comparaison entre les courbes continues et les courbes pointillées qui leur correspondent montre clairement que les distances au carré sont plus fidèles que leurs racines carrées malgré leur asymétrie beaucoup plus grande. Cette supériorité est d'environ trois points de corrélation (.03) et s'avère relativement constante dans les différentes conditions expérimentales.

Il est évident que cette différence de fidélité est attribuable à une différence dans la forme des distributions propres à $D^2$ et à $\sqrt{D^2}$, puisque l'une de ces mesures n'est qu'une transformation monotone de l'autre. Ce qui est moins évident est la raison pour laquelle la forme particulière de la distribution des $D^2$ - et spécialement sa plus grande asymétrie - favorise une plus grande fidélité[1].

[1] Nous ne croyons pas que cette supériorité soit une raison suffisante pour préférer les distances au carré aux distances simples. Comme Carroll (1961) le faisait remarquer, une corrélation linéaire peut se calculer sur n'importe quelles données mais son interprétation devient

(suite page 34)

Quant aux résultats essentiels de cette expérience, illustrés par les courbes continues de la figure 2, ils apparaissent, de façon générale, se conformer aux prédictions que nous avions faites. Aux deux niveaux de fidélité, les fonctions se croisent et, tant le point de départ de l'une (K = 2) que le point d'arrivée de l'autre (K = 1) correspondent assez étroitement aux valeurs prédites. L'examen de la Figure 2 est également instructif au niveau de la forme de ces fonctions. Les fonctions apparaissent linéaires dans le cas de la méthode 1 (celle où K = 1) et curvilinéaires dans le cas de la méthode 2. Ce qui s'avère étrange dans le cas de la méthode 2, c'est que la fidélité des distances y est d'abord une fonction décroissante, puis une fonction croissante du niveau de corrélation entre scores vrais. Ces fonctions non-monotones, nous apparaissent difficilement explicables au niveau théorique et pourraient bien être le fait des échantillons propres à cette expérience.

Une explication plausible serait que la corrélation échantillonnale entre scores vrais ait été légèrement négative dans les conditions expérimentales correspondant à une corrélation paramétrique nulle. Puisque les mêmes données ont servi à toutes les conditions expérimentales, ce biais négatif, toujours présent, aurait retardé l'obtention de corrélations échantillonnales supérieures à zéro. Ces dernières ne seraient apparues qu'à des niveaux de corrélations paramétriques bien supérieurs à zéro, d'où le retard observé dans la tendance croissante de ces fonctions. Cependant, le fait que cette tendance croissante apparaisse tardivement (autour d'une corrélation paramétrique de .50 entre scores vrais) conjugué au fait d'une grandeur échantillonnale respectable (n = 500), laisse à penser que notre interprétation pourrait bien ne pas être suffisante. Les

---

à ce point difficile dans les cas de non-normalité, qu'elle perd beaucoup de son utilité. Nous pensons que l'utilisateur de scores de distance a de bonnes raisons de préférer des valeurs de fidélité légèrement moins élevées, mais plus interprétables. Malgré ce dernier point de vue, les autres expériences de simulation réalisées dans cette étude ont utilisées les distances au carré. La raison n'est pas tant leur supériorité en terme de corrélation, que le fait que toutes les équations présentées dans les études précédentes traitaient de distances au carré. Les résultats de ces expériences seront ainsi plus facilement interprétables à la lumière de ces équations.

corrélations échantillonnales, qui auraient été nécessaires à sa vérification, n'ayent pas été recueillies, la façon la plus simple, à notre avis, de clarifier la nature réelle de ces fonctions, serait de répéter l'expérience sur de nouvelles données. Il serait probablement souhaitable que, contrairement à la présente expérience, des échantillons différents soient utilisés dans chacune des conditions expérimentales afin d'éviter les difficultés d'interprétations qu'entraînent une erreur échantillonnale constante.

Un dernier aspect, qu'il est intéressant d'analyser dans les résultats illustrés à la Figure 2, est le niveau où les fonctions, caractéristiques de chaque méthode, se croisent. L'examen des fonctions semble indiquer que ce niveau est plus élevé (autour de .45 de corrélation entre scores vrais) quand la fidélité est grande que quand elle est faible (autour de .20). Ces deux niveaux d'intersection(.45 et .20) ne sont sûrement que très approximatifs puisqu'ils proviennent de fonctions elles-mêmes imprécises. En effet, ces fonctions ne reposent que sur cinq valeurs lesquelles sont possiblement sujettes à une erreur échantillonnale constante - comme nous l'avons discuté précédemment. Malgré leur imprécision, ces deux points d'intersection semblent indiquer que le niveau de corrélation entre scores vrais, à partir duquel il devient avantageux de réunir les variables plutôt que de les séparer, se situe dans le voisinage du niveau de fidélité des variables. Contrairement aux valeurs observées ici, qui sont, dans un cas, inférieure à la fidélité et, dans l'autre cas, supérieure, il semble plus naturel que les vrais niveaux d'intersection soient ou bien inférieurs, ou bien égaux, ou bien supérieurs aux niveaux de fidélité des variables. Des données supplémentaires permettraient peut-être d'éclaircir ce dernier point.

## Expérience 2

### 1. Méthode

Une nouvelle expérience a été réalisée afin de clarifier certains aspects des résultats précédents, particulièrement la question de la forme des fonctions et celle du niveau où elles se croisent.

La même méthode de génération des scores vrais et observés a été utilisée ici. Par contre, trois modifications ont été apportées au plan expérimental. Premièrement, trois niveaux de fidélité (.20, .50 et .80) ont été étudiés au lieu de deux (.06 et .50), afin que les valeurs choisies soient plus représentatives de l'étendue possible des fidélités et aussi afin que trois points d'intersection, et non seulement deux, puissent être analysés. Deuxièmement, afin d'éviter la présence d'une erreur échantillonnale constante dans ces nouveaux résultats, des données différentes ont été obtenues pour chacune des 30 (et non plus 20) conditions expérimentales. En dernier lieu, la variable dépendante, c'est-à-dire la corrélation entre distances vraies et observées, a été calculée sur les distances au carré, $D^2$, et non sur les distances simples, de façon à faciliter la comparaison de ces résultats à ceux prédits par les équations des études précédentes, lesquelles portaient exclusivement sur les distances au carré.

## 2. Résultats

Le Tableau 2 présente les moyennes, sur dix échantillons, des corrélations entre distances vraies et distances observées propres à chaque condition expérimentale. Ces résultats sont illustrés à la Figure 3 sous la forme de six fonctions.

Il est clair que le fait de ne pas avoir utilisé les mêmes échantillons d'une condition à l'autre a considérablement nui à la clarté de ces fonctions. Ainsi, les formes linéaires ou curvilinéaires, observées précédemment, sont désormais obscurcies par le "bruit" échantillonnal. De plus, les niveaux d'intersection n'apparaissent plus liés aux niveaux de fidélité. Pour des fidélités respectives de .20, .50 et .80, nous obtenons ici des niveaux d'intersection d'environ .55, .30 et .75. Le seul aspect quelque peu positif à ces résultats est que le caractère non-monotone des fonctions, dans le cas de la méthode 2, qui avait été observé dans les résultats précédents, est ici contredit. En effet, pour chacune des trois fonctions obtenues, il y a augmentation de la fidélité entre le premier et le deuxième point.

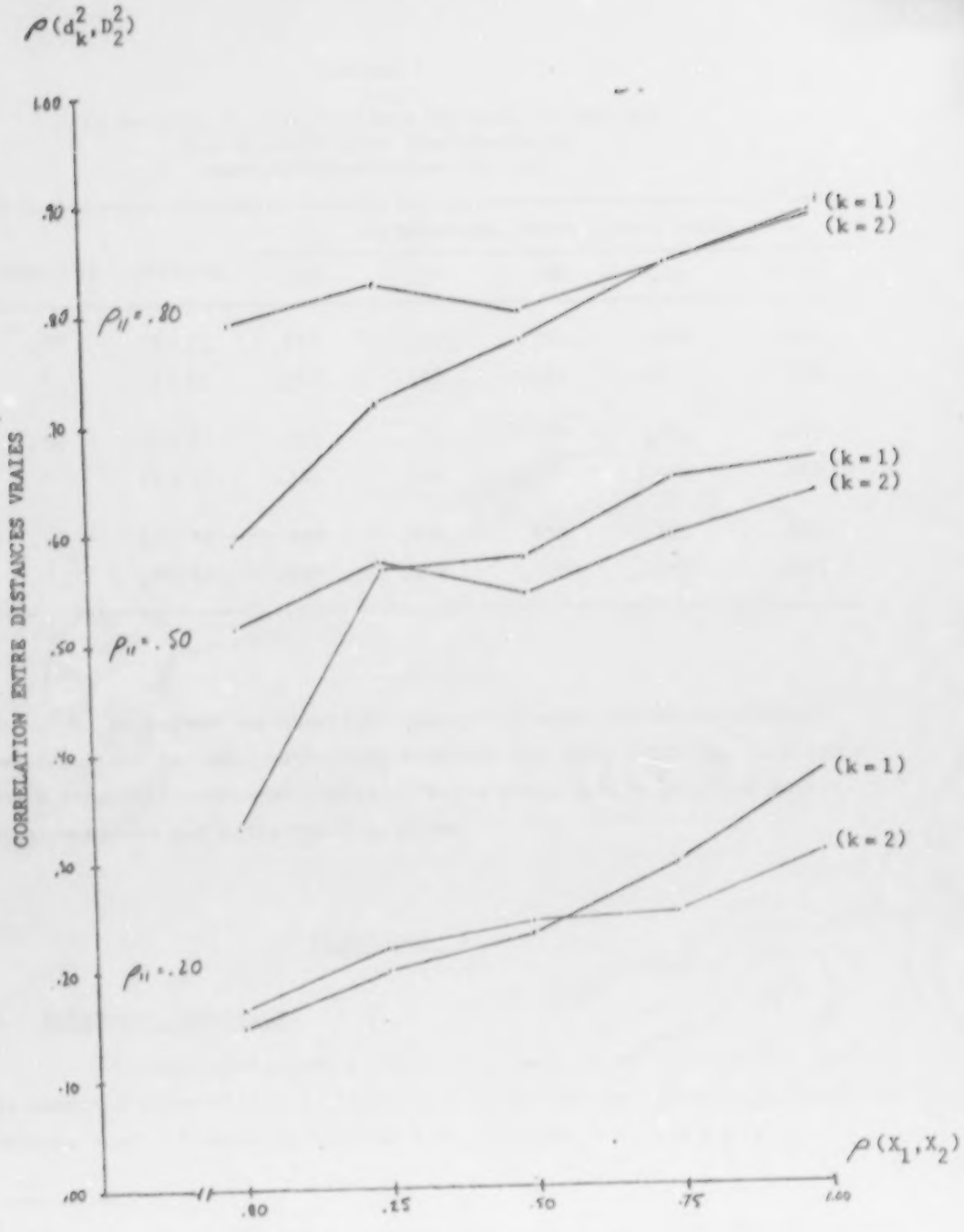



Figure 3. Résultats de l'expérience 2

Tableau 2

Expérience 2 - Corrélations moyennes de chacune des 30 conditions expérimentales pour 10 échantillons (n = 50)

| Fidélité | Méthode | corrélations entre scores vrais | | | | |
|---|---|---|---|---|---|---|
| | | .00 | .25 | .50 | .75 | 1.00 |
| .20 | (K = 2) | .165 | .220 | .241 | .250 | .303 |
| " | (K = 1) | .147 | .199 | .232 | .295 | .376 |
| .50 | (K = 2) | .512 | .571 | .539 | .594 | .632 |
| " | (K = 1) | .331 | .565 | .574 | .645 | .664 |
| .80 | (K = 2) | .789 | .825 | .800 | .841 | .884 |
| | (K = 1) | .589 | .717 | .772 | .842 | .889 |

En regard des questions auxquelles nous cherchions réponses, les résultats de cette expérience semblent par trop imprécis. Les grandeurs échantillonnales utilisées[1], toutes respectables qu'elles soient, n'en demeurent pas moins insuffisantes.

## Expérience 3

### 1. Méthode et résultats

La même expérience a été reprise mais en multipliant par trois le nombre d'échantillons à l'intérieur de chaque condition. Les résultats obtenus sont présentés au Tableau 3 et illustrés à la Figure 4.

---

[1] Les scores de 15,000 sujets avaient été créés à partir de 900,000 nombres aléatoires uniformes. Le temps total d'exécution a été de 120 secondes sur l'ordinateur CDC-6400 de l'Université du Québec, à Québec.

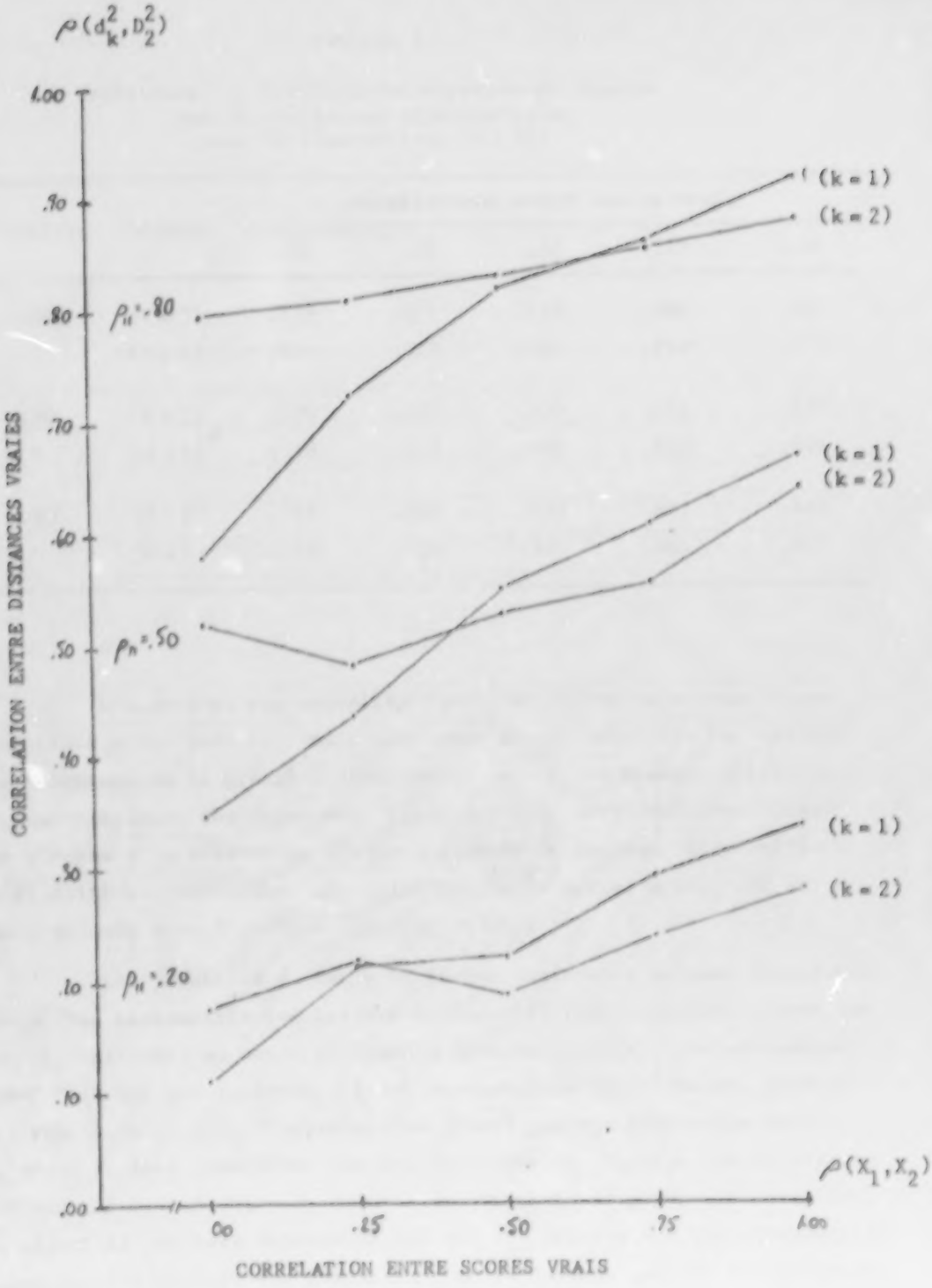



Figure 4. Résultats de l'expérience 3

Tableau 3

Expérience 3 - Corrélations moyennes de chacune des 30 conditions expérimentales pour 50 échantillons (n = 50)

| Fidélité | Méthode | corrélations entre scores vrais | | | | |
|---|---|---|---|---|---|---|
| | | .00 | .25 | .50 | .75 | 1.00 |
| .20 | (K = 2) | .172 | .221 | .190 | .240 | .281 |
| " | (K = 1) | .114 | .215 | .220 | .294 | .337 |
| .50 | (K = 2) | .520 | .485 | .529 | .556 | .639 |
| " | (K = 1) | .348 | .437 | .553 | .610 | .669 |
| .80 | (K = 2) | .796 | .810 | .832 | .855 | .881 |
| " | (K = 1) | .580 | .726 | .821 | .861 | .887 |

L'examen de ces nouvelles fonctions permet de constater que leur tracé s'est amélioré, mais sans pour autant atteindre la netteté des fonctions de la première expérience. Ainsi, il demeure difficile de leur attribuer une forme soit linéaire, soit curvilinéaire. Quant aux niveaux d'intersection, ils apparaissent de nouveau, être reliés à la fidélité des variables. Pour des fidélités égales à .20, .50 et .80, leurs valeurs sont d'environ .25, .40 et .65.

Les résultats de cette troisième expérience ne sont finalement guère plus instructifs que les précédents. Il faut cependant avouer que tenter d'estimer, au moyen de données échantillonnales, des phénomènes aussi délicats que la forme précise de certaines fonctions et, particulièrement, leur point d'intersection, n'est pas une entreprise facile. De plus, il faut considérer que ces fonctions ne traitent pas de mesures individuelles, mais bien de mesures descriptives de groupes d'individus. En effet, la variable dépendante qui est ici étudiée est une corrélation, laquelle statistique, par définition, ne se calcule que sur des échantillons de mesures. Ce genre d'étude, compte tenu du nombre immense de don-

nées individuelles qu'il nécessite, semble donc exiger une stratégie expérimentale mûrement réfléchie, afin de contrevenir au "bruit" échantillonnal.

## Expérience 4

### 1. Méthode

L'expérience précédente a été recommencée mais en répétant les mêmes échantillons dans toutes les conditions expérimentales. Comme en témoignent les courbes de la première expérience, cette procédure a l'heur d'éliminer presque entièrement toute fluctuation échantillonnale. Comme nous le notions, elle présente aussi le désavantage de reproduire tout biais échantillonnal à travers l'ensemble des conditions expérimentales. Ce biais peut cependant être considérablement amenuisé en utilisant des échantillons de grandeur suffisante. Tout compte fait, il apparaît encore préférable à cette autre difficulté, rencontrée dans les expériences 2 et 3, qu'entraînent les fluctuations échantillonnales. Dans le but de restreindre le plus possible l'influence de ce biais, le nombre d'échantillons a de nouveau été augmenté, soit de 30 à 50.

Afin d'améliorer le tracé des fonctions, le nombre de niveaux de corrélation entre scores vrais a également été augmenté de cinq à neuf. Ainsi, en plus des valeurs de .00, .25, .50, .75 et 1.00, les valeurs intermédiaires de .125, .375, .625 et .875 ont été étudiées. L'Appendice B reproduit le programme principal utilisé dans cette expérience.

### 2. Résultats

Les corrélations moyennes résultant de chacune des 54 conditions expérimentales sont rapportées au Tableau 4 et les fonctions auxquelles elles donnent lieu sont représentées à la Figure 5.

L'examen de cette figure est instructif à plus d'un titre. Premièrement, les fonctions obtenues ici, compte tenu de leur régularité presque parfaite, confirment la forme linéaire de la relation dans le cas

nées individuelles qu'il nécessite, semble donc exiger une stratégie expérimentale mûrement réfléchie, afin de contrevenir au "bruit" échantillonnal.

## Expérience 4

### 1. Méthode

L'expérience précédente a été recommencée mais en répétant les mêmes échantillons dans toutes les conditions expérimentales. Comme en témoignent les courbes de la première expérience, cette procédure a l'heur d'éliminer presque entièrement toute fluctuation échantillonnale. Comme nous le notions, elle présente aussi le désavantage de reproduire tout biais échantillonnal à travers l'ensemble des conditions expérimentales. Ce biais peut cependant être considérablement amenuisé en utilisant des échantillons de grandeur suffisante. Tout compte fait, il apparait encore préférable à cette autre difficulté, rencontrée dans les expériences 2 et 3, qu'entraînent les fluctuations échantillonnales. Dans le but de restreindre le plus possible l'influence de ce biais, le nombre d'échantillons a de nouveau été augmenté, soit de 30 à 50.

Afin d'améliorer le tracé des fonctions, le nombre de niveaux de corrélation entre scores vrais a également été augmenté de cinq à neuf. Ainsi, en plus des valeurs de .00, .25, .50, .75 et 1.00, les valeurs intermédiaires de .125, .375, .625 et .875 ont été étudiées. L'Appendice B reproduit le programme principal utilisé dans cette expérience.

### 2. Résultats

Les corrélations moyennes résultant de chacune des 54 conditions expérimentales sont rapportées au Tableau 4 et les fonctions auxquelles elles donnent lieu sont représentées à la Figure 5.

L'examen de cette figure est instructif à plus d'un titre. Premièrement, les fonctions obtenues ici, compte tenu de leur régularité presque parfaite, confirment la forme linéaire de la relation dans le cas

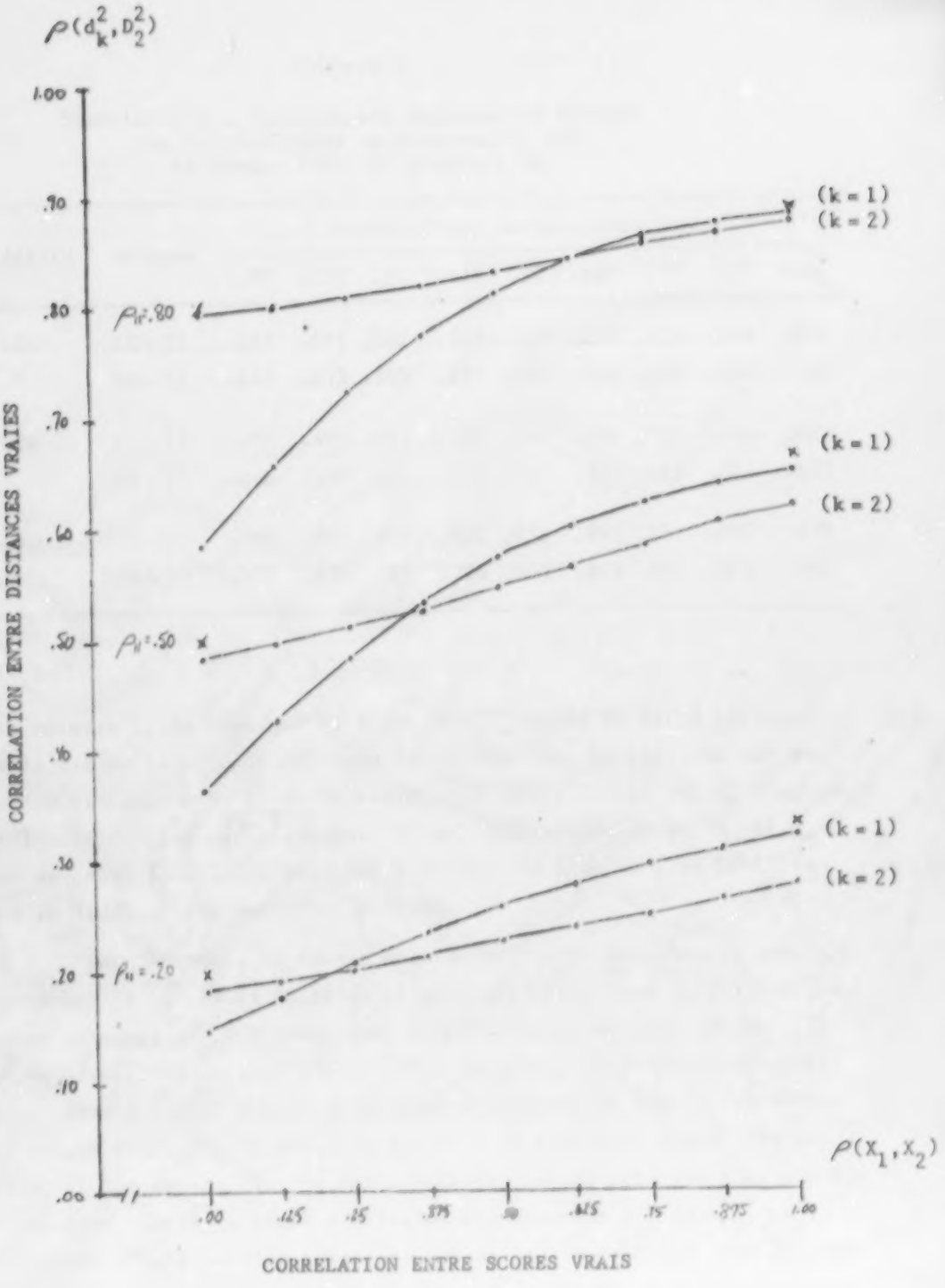



Figure 5. Résultats de l'expérience 4

Tableau 4

Expérience 4 - Corrélations moyennes de chacune des 54 conditions expérimentales sur 50 échantillons de grandeur 50

| Fidélité | Méthode | corrélations entre scores vrais | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | .00 | .125 | .25 | .375 | .50 | .625 | .75 | .875 | 1.00 |
| .20 | (K = 2) | .183 | .191 | .202 | .214 | .227 | .239 | .252 | .266 | .279 |
| " | (K = 1) | .147 | .179 | .209 | .237 | .260 | .280 | .297 | .310 | .322 |
| .50 | (K = 2) | .486 | .498 | .513 | .529 | .547 | .566 | .585 | .604 | .622 |
| | (K = 1) | .366 | .430 | .486 | .534 | .572 | .602 | .625 | .641 | .652 |
| .80 | (K = 2) | .795 | .801 | .810 | .820 | .832 | .843 | .855 | .867 | .878 |
| | (K = 1) | .585 | .659 | .723 | .774 | .814 | .843 | .863 | .876 | .882 |

de la méthode 2, de même que la forme curvilinéaire de cette relation dans le cas de la méthode 1. Dans ce dernier cas, la fidélité est une fonction croissante du degré de corrélation entre scores vrais, fonction dont l'accélération est négative. Il est intéressant de noter que la forme de cette fonction a tendance à devenir de plus en plus linéaire quand la fidélité des variables diminue.

Deuxièmement, le niveau d'intersection de ces courbes apparait clairement lié au niveau de fidélité des variables. Les points d'intersection observés graphiquement sont respectivement de .21, .35 et .625 pour des fidélités de .20, .50 et .80. Le niveau de corrélation entre scores vrais à partir duquel il devient avantageux de réunir les variables serait donc plus ou moins supérieur à la fidélité, quand elle est faible, et lui serait plus ou moins inférieur, quand elle est plus élevée. Une relation similaire avait d'ailleurs été observée à l'expérience 1: des niveaux d'intersection voisinant .20 et .45 avaient été obtenus pour des fidélités respectives de .06 et .50. Nous avions alors émis l'opinion que les niveaux d'intersection devraient être ou inférieurs ou égaux

ou supérieurs à la fidélité. Les présents résultats semblent plutôt indiquer le contraire.

Troisièmement, on y remarque que le gain de fidélité qu'entraîne la réunion ou la séparation des variables, selon le cas, n'est pas également avantageux à tous les niveaux de fidélité des variables. Considérons, à titre d'exemple, les deux cas extrêmes où les scores vrais sont en corrélation soit nulle, soit parfaite. Dans le premier ($\rho(X_1,X_2) = 0$), on observe que le gain de fidélité réalisé en séparant les variables est d'autant plus grand que la fidélité des variables est élevée. Dans le second ($\rho(X_1,X_2) = 1.0$), c'est le contraire qui se produit: le gain de fidélité réalisé en réunissant les variables est d'autant plus faible que la fidélité est élevée. Ces conclusions, bien qu'elles n'aient pas été mentionnées, étaient déjà évidentes dans les résultats illustrés à la Figure 1 de l'étude 1. De façon plus générale, l'examen, dans leur ensemble, des fonctions de la Figure 5, indique que la question de réunir ou de séparer les variables dans le calcul des distances ne se pose pas avec la même acuité à tous les niveaux de fidélité. Dans le cas d'une fidélité élevée, l'écart maximum entre les deux fonctions peut être très considérable. Par contre, dans le cas d'une fidélité faible, l'écart maximum est beaucoup moindre. Il s'en suit que c'est surtout dans le cas où les variables sont très fidèles, que le choix de la méthode de calculer les distances s'avère important.

Une autre conclusion d'intérêt pratique qui ressort de l'examen de ces fonctions, est que la méthode 2 est presque toujours préférable quand les variables sont très fidèles, et la méthode 1, quand elles le sont très peu. Ainsi pour une fidélité de .80, la méthode 2 est très supérieure dans la majorité des cas et, très peu inférieure, dans le reste des cas. Il s'en suit qu'il y a beaucoup à gagner, et peu à perdre, à la choisir dans le cas de fidélités élevées. C'est, jusqu'à un certain point, le contraire qui se produit dans le cas de faibles fidélités. La méthode 1 s'avère presque toujours le meilleur choix. Cependant, le gain de fidélité qu'elle entraîne y est moins grand alors que la perte de fidélité qu'elle entraîne, dans la minorité des cas où elle est moins avantageuse, y est plus grand.

Un dernier point à cette discussion concerne la présence ou l'absence relative d'un biais échantillonnal dans les résultats obtenus. Il est évidemment difficile de savoir si les fonctions obtenues correspondent étroitement aux fonctions "réelles" puisque ces dernières nous sont inconnues. Cependant, un début de réponse est peut-être possible puisque les équations 12 et 17, développées à l'étude 2, permettent de prédire soit la valeur initiale, soit la valeur terminale, de chacune de ces six fonctions. Ces six valeurs sont représentées par des 'X' à la Figure 5. On remarque que, dans tous ces cas, les corrélations prédites sont très légèrement supérieures aux moyennes des corrélations observées. Nous ne connaissons pas de test pour vérifier si cet écart est significatif. En supposant qu'il le soit, c'est-à-dire que sa grandeur dépasse ce que nous serions en droit d'attendre du hasard échantillonnal, une explication, autre que celle du hasard, pourrait peut-être être avancée. En effet, il n'est pas impossible que l'estimation d'une corrélation paramétrique au moyen de la moyenne arithmétique de plusieurs corrélations échantillonnales produise un estimé biaisé du paramètre d'intérêt, et cela même, dans un cas comme le nôtre, où les transformations z de Fisher ont été utilisées. Que le simple hasard échantillonnal, ou toute autre raison, soit la cause de cet écart systématique, son importance n'en demeure pas moins très réduite dans les présents résultats. Aussi, croyons-nous justifié de conclure que ces fonctions, qui ont été obtenues par voie de simulation et à partir desquelles un certain nombre de conclusions ont été formulées, correspondent vraisemblablement de façon très étroite aux fonctions réelles. Cette correspondance pourrait même être considérée comme presque parfaite, à la condition de tenir compte de l'écart systématique, qui a été observé et de considérer ces fonctions comme caractérisant des niveaux de fidélité, non pas de .20, .50 et .80, mais plus précisément de .19, .49 et .79.

## Résumé et conclusion

En guise de prolongement à l'étude 1, la fidélité des distances, $D^2$, a été étudiée dans les cas intermédiaires où les scores vrais des

variables entretiennent un niveau quelconque de corrélation. Quatre expériences sur des données simulées ont été rapportées. Les trois premières se sont finalement avérées plus instructives au niveau de la stratégie expérimentale que nécessite ce genre d'étude qu'au niveau des questions auxquelles nous cherchions réponses. Par contre, la dernière expérience a permis de produire des résultats fort détaillés. La conclusion essentielle à cette étude est que la question de réunir ou de séparer les variables se pose surtout dans le cas de fidélités intermédiaires. Il est presque toujours avantageux de séparer les variables quand leur fidélité est très élevée, et de les réunir dans le cas contraire. Un prolongement souhaitable à cette étude serait de développer les équations permettant d'expliquer les résultats obtenus.

## ETUDE 6

### La fidélité des distances calculées sur plusieurs variables normales quelconques

La présente étude développe la formule de la fidélité des distances dans le cas général de plusieurs variables normales reliées.

Elle constitue un achèvement à l'étude 2 où seules les fidélités, dans les cas plus simples de une ou plusieurs variables indépendantes, avaient été développées. Si l'étude 2 peut ici être complétée, c'est grâce au développement, dans l'étude 4, de la formule de la covariance entre valeurs au carré, $\sigma(X^2,Y^2)$. Les précisions apportées par l'étude 3 sont également incorporées aux présents développements: l'indépendance expérimentale entre scores vrai et d'erreur y est postulée et la fidélité y est définie comme le carré de la corrélation avec les distances vraies. En dernier lieu, ces nouvelles formules permettent d'expliquer analytiquement une partie des résultats de l'étude 5 et constituent, ainsi, le prolongement souhaité à cette dernière étude.

### 1. Moyenne et variance

Soit k variables aléatoires normales, $X_1$, $X_2$,...$X_i$...$X_k$, dont les moyennes, les variances et les covariances sont quelconques.

La distance, sur ces k variables, entre deux individus j et j' se symbolise $D^2_{k_{jj'}}$ et s'obtient en sommant les carrés des différences entre leurs scores respectifs sur chacune des variables, soit

$$D^2_{k_{jj'}} = \sum_{i=1}^{k} (X_{ij} - X_{ij'})^2$$

ou, plus simplement,

$$D^2_k = \sum_{i=1}^{k} (X_i - X_{i'}) = \sum_{i=1}^{k} D_i^2 \qquad (2.13)$$

où $D_i$ représente la différence algébrique entre les scores de ces deux individus sur la variable i.

La distance $D_k^2$, à l'instar des k variables dont elle origine, est aussi une variable aléatoire. Elle est définie, non pas sur l'ensemble des individus de la population, mais bien sur l'ensemble de toutes les paires possibles entre les individus de cette population. Sa distribution, contrairement à celle des variables d'origine, n'est cependant pas normale. Dans le cas particulier où les k variables seraient indépendantes et auraient des variances égales, la forme de la distribution de $D_k^2$ serait exactement celle du chi carré avec k degrés de liberté.

La moyenne de la distribution de $D_k^2$ s'obtient par

$$\mathcal{E}(D_k^2) = \mathcal{E}(\sum_i^k D_i^2) = \sum_i^k (\mathcal{E} D_i^2)$$

$$= 2 \sum^k \sigma^2(X_i) \qquad (2.14)$$

puisque $\mathcal{E} D_i^2 = 2\sigma^2(X_i)$ par (4.37)[1].

Quant à la variance de $D_k^2$, elle s'obtient par la formule de la variance d'une somme, soit

$$\sigma^2(D_k^2) = \sigma^2(\sum_i^k D_i^2)$$

$$= \sum_{ij}^{kk} \sigma(D_i^2, D_j^2)$$

$$= \sum_i^k \sigma^2(D_i^2) + \sum_{i \neq j}^{kk} \sigma(D_i^2, D_j^2)$$

$$= 8 \left[ \sum_i^k \sigma^4(X_i) + \sum_{i \neq j}^{kk} \sigma^2(X_i, X_j) \right] \qquad (6.1)$$

puisque $\sigma^2(D_i^2) = \sigma^4(X_i)$, par (4.40) et $\sigma(D_i^2, D_j^2) = \sigma^2(X_i, X_j)$, par (4.41).

## 2. Fidélité

Soit les scores observés, $x_i$, obtenus sur k variables, chacun

---

[1] Les équations seront désormais symbolisées par deux nombres entre parenthèses: le premier réfère au numéro de l'étude et le second au numéro de l'équation à l'intérieur de l'étude.

étant défini comme la somme d'un score vrai, $X_i$, et d'un score d'erreur, $E_i$,

$$x_i = X_i + E_i \quad \text{pour tout } i,\ i = 1, 2 \ldots k \tag{6.2}$$

tels que $X_i$ et $E_i$ sont expérimentalement indépendants et que $\mathcal{E}(E_i) = 0$, pour tout i.

La différence entre deux valeurs de $x_i$ échantillonnées au hasard peut aussi être définie comme la somme d'une composante vraie et d'une composante d'erreur, soit

$$\begin{aligned} d_i &= x_i - x'_i \\ &= X_i + E_i - X'_i - E'_i \\ &= D_i - Ed_i \end{aligned} \tag{6.3}$$

où $D_i = X_i - X'_i$ et $Ed_i = E_i - E'_i$. Compte tenu que X et E ainsi que X' et E' sont expérimentalement indépendants et compte tenu que x et x' sont aussi expérimentalement indépendants - puisque échantillonnés au hasard - il s'en suit, par transitivité, que D et Ed sont également expérimentalement indépendants.

Si les distances observées sont définies, par convention, comme les distances calculées sur les scores observés $x_i$, i.e.

$$d_k^2 = \sum_i^k (x_i - x'_i)^2 = \sum_i^k d_i^2 \tag{6.4}$$

et les distances vraies comme celles calculées sur les scores vrais $X_i$, i.e.

$$D_k^2 = \sum_i^k (X_i - X'_i)^2 = \sum_i^k D_i^2 \tag{6.5}$$

la fidélité des distances pourra, à son tour, être définie comme le carré de la corrélation entre distances vraies et observées, i.e. $\rho^2(d_k^2, D_k^2)$. or

$$\begin{aligned} \rho^2(d_k^2, D_k^2) &= \rho^2(\Sigma d_i^2, \Sigma D_i^2) \\ &= \sigma^2(\Sigma d_i^2, \Sigma D_i^2) \,/\, \sigma^2(\Sigma d_i)\ \sigma^2(\Sigma D_i^2) \end{aligned} \tag{6.6}$$

Le numérateur de (6.6) est lui-même égal au carré de

$$\sigma\left(\sum_{i}^{k} d_i^2,\ \sum_{j}^{k} D_j^2\right) = \sum_{ij}^{kk} \sigma(d_i^2,\ D_j^2)$$

$$= \sum_{i}^{k} \sigma(d_i^2,\ D_i^2) + \sum_{i \neq j}^{kk} \sigma(d_i^2,\ D_j^2) \qquad (6.7)$$

Or, il a déjà été démontré que la covariance entre distances vraies et observées sur une même variable est égale à la variance des distances vraies.

$$\sigma(d_i^2,\ D_i^2) = \sigma^2(D_i^2) \qquad (3.4)$$

Quant à la covariance entre distances vraies sur une variable et distances observées sur une autre, elle est égale à

$$\sigma(d_i^2,\ D_j^2) = \sigma(D_i^2 + Ed_i + 2D_i Ed_i,\ D_j^2)$$

$$= \sigma(D_i^2, D_j^2) + \sigma(Ed_i, D_j^2) + 2\sigma(D_i Ed_i, D_j^2)$$

où $\qquad \sigma(Ed_i,\ D_j^2) = 0 \qquad (6.8)$

suite à l'indépendance expérimentale entre $D_i$ et $Ed_i$ laquelle implique, a fortiori, celle entre $D_j$ et $Ed_i$, et où

$$\sigma(D_i Ed_i, D_j^2) = \mathcal{E} D_j^2 D_i Ed_i - \mathcal{E} D_j^2\ \mathcal{E} D_i Ed_i$$

$$= \mathcal{E} D_j^2 D_i \mathcal{E} Ed_i - \mathcal{E} D_j^2\ \mathcal{E} D_i \mathcal{E} Ed_i = 0 \qquad (6.9)$$

puisque $\mathcal{E} Ed_i = \mathcal{E} E_i - \mathcal{E} E_i' = 0 - 0 = 0$. Ces deux dernières covariances étant nulles, (6.8) et (6.9), il s'en suit que

$$\sigma(d_i^2,\ D_j^2) = \sigma(D_i^2,\ D_j^2) \qquad (6.10)$$

Les résultats (3.4) et (6.10) permettent donc d'exprimer la covariance entre distances vraies et observées sur k variables, soit $D_k^2$ et $d_k^2$, comme la variance des distances vraies.

Ainsi

$$\sigma(\sum_{i}^{k} d_i^2, \sum_{j}^{k} D_j^2) = \sum_{i}^{k} \sigma(d_i^2, D_i^2) + \sum_{i \neq j}^{kk} \sigma(d_i^2, D_j^2) \qquad (6.7)$$

$$= \sum_{i}^{k} \sigma^2(D_i^2) + \sum_{i \neq j}^{kk} \sigma(D_i^2, D_j^2)$$

$$= \sigma^2(\sum_{i}^{k} D_i^2) = \sigma^2(D_k^2) \qquad (6.11)$$

Ce dernier résultat permet, à son tour, d'exprimer la corrélation au carré entre distances vraies et observées comme le rapport de leurs variances respectives. Ainsi

$$\rho^2(d_k^2, D_k^2) = \sigma^2(\Sigma d_i^2, \Sigma D_i^2) \,/\, \sigma^2(\Sigma d_i^2)\, \sigma^2(\Sigma D_i^2) \qquad (6.6)$$

$$= \sigma^4(D_k^2) \,/\, \sigma^2(d_k^2)\, \sigma^2(D_k^2)$$

$$= \sigma^2(D_k^2) \,/\, \sigma^2(d_k^2) \qquad (6.12)$$

Comme nous l'avons discuté ailleurs (voir étude 3), le résultat précédent, (6.12) ne postule rien d'autre que l'indépendance expérimentale de x et x' et celle de X et E. Par contre, si l'on veut exprimer la fidélité des distances, $d_k^2$, en fonction des deux premiers moments de la distribution des k variables initiales, il devient nécessaire de postuler leur normalité. Pour ce faire, il suffit d'utiliser (6.1) dans la solution du numérateur et du dénominateur de (6.12). Ainsi

$$\rho^2(d_k^2, D_k^2) = \frac{8\left[\sum_{i}^{k} \sigma^4(X_i) + \sum_{i \neq j}^{kk} \sigma^2(X_i, X_j)\right]}{8\left[\sum_{i}^{k} \sigma^4(x_i) + \sum_{i \neq j}^{kk} \sigma^2(x_i, x_j)\right]}$$

$$= \frac{\sum_{i}^{k} \sigma^4(X_i) + \sum_{i \neq j}^{kk} \sigma^2(X_i, X_j)}{\sum_{i}^{k} \sigma^4(x_i) + \sum_{i \neq j}^{kk} \sigma^2(X_i, X_j)} \qquad (6.13)$$

puisque la covariance entre scores observés est toujours égale à celle entre scores vrais.

Dans le cas particulier où les k scores sont <u>parallèles</u>[1] (i.e. identité des scores vrais et égalité des variances d'erreur), il est possible de simplifier (6.13) puisque toutes les variances (vraies ou observées) sont égales et puisque les covariances sont égales aux variances vraies. Ainsi

$$\rho^2(d_k^2, D_k^2) = \frac{k^2\sigma^4(X_i)}{k\sigma^4(x_i) + k(k-1)\sigma^4(X_i)} \qquad (6.14)$$

Mais puisque il est connu que $\sigma^2(X_i) = \rho^2(x_i, X_i)\,\sigma^2(x_i)$ et que, conséquemment, $\sigma^4(x_i) = \sigma^4(X_i)/\rho^4(x_i, X_i)$, il s'en suit que

$$\rho^2(d_k^2, D_k^2) = \frac{k^2\sigma^4(X_i)}{\frac{k\sigma^4(X_i)}{\rho^4(x_i, X_i)} + k(k-1)\sigma^4(X_i)}$$

$$= \frac{k\,\rho^4(x_i, X_i)}{1 + (k-1)\rho^4(x_i, X_i)} \qquad (6.15)$$

Mais puisque, par (2.12), $\rho^4(x_i, X_i) = \rho^2(d_i^2, D_i^2)$, la fidélité des distances sur k mesures parallèles peut aussi être exprimée en fonction de la fidélité des distances sur une seule mesure, soit

$$\rho^2(d_k^2, D_k^2) = \frac{k\,\rho^2(d_i^2, D_i^2)}{1 + (k-1)\rho^2(d_i^2, D_i^2)} \qquad (6.16)$$

Ce dernier résultat démontre que la formule bien connue de Spearman-Brown s'applique tout aussi bien aux scores de distances.

---

[1] Si des mesures sont parallèles et si, de plus, leurs scores vrai et d'erreur sont expérimentalement indépendants - comme nous avons été amené à le postuler ici - il s'en suit que ces mesures peuvent alors être définies comme <u>équivalentes</u>, conformément à la définition présentée par Lord et Novick (1968, p. 46).

## ETUDE 7

### La fidélité et la validité des distances calculées après sommation de plusieurs variables quelconques

Dans les expériences de simulation précédentes (études 1 et 5), deux types de "distances" avaient été comparés selon que, préalablement à leur calcul, les k variables étaient séparées ou réunies.

Si les "distances observées" sont définies comme celles calculées sur les scores observés $x_i$ et $x'_i$, et les "distances vraies", celles calculées sur les scores vrais $X_i$ et $X'_i$, les distances vraies et observées de l'un et l'autre type s'obtiendront par

$$D_k^2 = \sum_{i=1}^{k} (X_i - X'_i)^2 = \sum_{i=1}^{k} D_i^2 \qquad (6.3)$$

$$d_k^2 = \sum_{i=1}^{k} (x_i - x'_i)^2 = \sum_{i=1}^{k} d_i^2 \qquad (6.4)$$

$$D_1^2 = (\sum_{i=1}^{k} X_i - \sum_{i=1}^{k} X'_i)^2 = (\sum_{i=1}^{k} D_i)^2 \qquad (7.1)$$

$$d_1^2 = (\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} x'_i)^2 = (\sum_{i=1}^{k} d_i)^2 \qquad (7.2)$$

où majuscules et minuscules représentent respectivement le niveau vrai et le niveau observé, et où les indices "1" et "k" indiquent si les distances proviennent d'une seule variable après sommation ou, au contraire, des k variables.

Dans ces expériences, la corrélation entre $d_k^2$ et $D_k^2$ avait été comparée à celle entre $d_1^2$ et $D_k^2$ à divers niveaux de fidélité et à divers niveaux de corrélation entre scores vrais. Ces deux corrélations avec les distances vraies $D_k^2$ avaient alors été considérées comme les fidélités respectives de ces deux distances observées $d_1^2$ et $d_k^2$. L'étude 6 a permis

de produire la formule de la première de ces corrélations, soit $\rho(d_k^2, D_k^2)$, dans le cas général de plusieurs variables normales dont les moyennes, les variances, les fidélités et les covariances sont quelconques. Quant à la présente étude, elle constitue un prolongement naturel de cette dernière, puisque les autres corrélations entre les quatre distances $d_k^2$, $d_l^2$, $D_k^2$ et $D_l^2$, prises deux à deux, y sont développées. Il est à noter que la fidélité des distances $d_l^2$ sera désormais définie comme la corrélation $\rho(d_l^2, D_l^2)$ et non plus comme la corrélation $\rho(D_k^2, d_l^2)$. Quant à cette dernière, elle pourrait être définie comme une validité en autant que la variable qu'on cherche à estimer corresponde aux distances $D_k^2$. Un résumé des formules les plus importantes des études 6 et 7 est présenté sous la forme du Tableau 7.1.

1. Les variances

La variance de $D_k^2$, comme il l'a été démontré précédemment, est égale à

$$\sigma^2(D_k^2) = 8\left[\sum_i^k \sigma^4(X_i) + \sum_{i\neq j}^{k\,k} \Sigma\sigma^2(X_i, X_j)\right] \tag{6.1}$$

$$= 8 \sum_{ij}^{kk} \Sigma\Sigma\sigma^2(X_i, X_j) \tag{7.3}$$

Celle de $d_k^2$ s'obtient de façon similaire à (6.1) et elle est égale à

$$\sigma^2(d_k^2) = 8 \sum_{ij}^{kk} \Sigma\Sigma\sigma^2(x_i, x_j) \tag{7.4}$$

Quant aux deux autres variances, elles s'obtiennent par

$$\sigma^2(D_l^2) = \sigma^2\left[\left(\sum_i^k D_i\right)^2\right]$$

$$= 2\left[\sigma^2\left(\sum_i^k D_i\right)\right]^2$$

puisque, par (4.26), $\sigma^2(X^2) = 2\left[\sigma^2(X)\right]^2$,

$$= 2\left[\sum_{ij}^{kk} \sigma(D_i, D_j)\right]^2$$

$$= 2\left[\sum_{ij}^{kk} 2\sigma(X_i, X_j)\right]^2$$

$$= 8\left[\sum_{ij}^{kk} \sigma(X_i, X_j)\right]^2 \qquad (7.5)$$

puisque, par (4.36), $\sigma(D_x, D_y) = 2\sigma(X, Y)$; et, de façon similaire,

$$\sigma^2(d_1^2) = 8\left[\sum_{ij}^{kk} \sigma(x_i, x_j)\right]^2 \qquad (7.6)$$

2. Les fidélités

Il a été démontré à l'étude 6 que

$$\sigma(D_k^2, d_k^2) = \sigma^2(D_k^2) \qquad (6.11)$$

et conséquemment la fidélité des $d_k^2$ était égale à

$$\rho^2(D_k^2, d_k^2) = \sigma^2(D_k^2) \,/\, \sigma^2(d_k^2) \qquad (6.12)$$

$$= \sum_{ij}^{kk} \sigma^2(X_i, X_j) \,/\, \sum_{ij}^{kk} \sigma^2(x_i, x_j)$$

$$= \frac{\sum_{i}^{k} \sigma^4(X_i) + \sum_{i\neq j}^{kk} \sigma^2(X_i, X_j)}{\sum_{i}^{k} \sigma^4(x_i) + \sum_{i\neq j}^{kk} \sigma^2(X_i, X_j)} \qquad (6.13)$$

puisque la covariance entre scores observés est égale à la covariance entre scores vrais.

Dans le cas particulier où les k variables partagent les mêmes variances vraies ($\sigma_x^2$), les mêmes fidélités ($\rho_{11}$) et les mêmes intercorrélations ($\rho_{12}$), c'est-à-dire dans le cas de variables <u>comparables</u> comme nous les appellerons à l'avenir, l'équation (6.13) se simplifie en

$$\rho^2(D_k^2, d_k^2) = \frac{1 + (k-1)\rho_{12}^2}{1/\rho_{11}^2 + (k-1)\rho_{12}^2} \qquad (7.7)$$

Dans le cas encore plus particulier où, de plus, les covariances sont égales aux variances vraies, c'est-à-dire le cas de variables <u>parallèles</u> comme définis par Lord et Novick (1968, pp. 58-59), l'équation (7.7) se simplifie à son tour en

$$\rho^2(D_k^2,\ d_k^2) = \frac{k\rho_{11}^2}{1 + (k-1)\rho_{11}^2} \tag{6.15}$$

$$= \frac{k\rho^2(D_i^2,\ d_i^2)}{1 + (k-1)\rho^2(D_i^2,\ d_i^2)} \tag{6.16}$$

puisque, par (2.12), $\rho_{11}^2 = \rho^4(X_i,\ x_i) = \rho^2(D_i^2,\ d_i^2)$. Ce dernier résultat montre, comme nous l'avions noté à l'étude précédente, que la formule bien connue de Spearman-Brown s'applique tout aussi bien aux scores de distances $d_k^2$.

Quant aux distances calculées après sommation des k variables, $d_i^2$, leur covariance avec les distances vraies correspondantes, $D_i^2$, s'obtient par

$$\sigma(D_i^2,\ d_i^2) = \sigma\left[(\sum_i^k D_i)^2,\ (\sum_i^k d_i)^2\right]$$

$$= 2\left[\sigma(\sum_i^k D_i,\ \sum_i^k d_i)^2\right]$$

puisque, par (4.27), $\sigma(U_X^2,\ U_y^2) = 2\sigma^2(X,\ Y)$

$$= 2\left[\sum_{ij}^{kk} \sigma(D_i,\ d_j)\right]^2$$

$$= 2\left[\sum_{ij}^{kk} \sigma(D_i,\ D_j)\right]^2$$

puisque, suite à la définition (3.3), $\sigma(D_i,\ d_j) = \sigma(D_i,\ D_j)$ pour tout i et tout j,

$$= \sigma^2(D_i^2) \tag{7.8}$$

puisque, dans le développement de (7.5), il a été démontré que

$$\sigma^2(D_1^2) = 2\left[\sum_i^k \sum_j^k \sigma(D_i, D_j)\right]^2.$$

La conséquence de ce dernier résultat est que la fidélité des distances $d_1^2$ sera aussi égale au rapport de la variance vraie sur la variance observée, soit

$$\rho^2(D_1^2, d_1^2) = \sigma^2(D_1^2, d_1^2) \,/\, \sigma^2(D_1^2)\,\sigma^2(d_1^2)$$

$$= \sigma^4(D_1^2) \,/\, \sigma^2(D_1^2)\,\sigma^2(d_1^2)$$

$$= \sigma^2(D_1^2) \,/\, \sigma^2(d_1^2) \qquad (7.9)$$

$$= \left[\frac{\sum_i^k \sum_j^k E\,\sigma(X_i, X_j)}{\sum_i^k \sum_j^k \sigma(x_i, x_j)}\right]^2$$

$$= \left[\sigma^2\left(\sum_i^k X_i\right) \,/\, \sigma^2\left(\sum_i^k x_i\right)\right]^2$$

$$= \left[\rho^2\left(\sum_i^k X_i, \sum_i^k x_i\right)\right]^2 \qquad (7.10)$$

Ce résultat est le même que celui de l'équation (2.12) où il était question de la fidélité des distances obtenues à partir d'une seule variable. La seule différence entre ces deux équations est que la variable considérée ici est une "variable composée".

Dans le cas particulier de variables comparables - telles que définies précédemment - la fidélité devient alors égale à

$$\rho^2(D_1^2, d_1^2) = \left[\frac{\sum_i^k \sum_i^k \sigma(X_i, X_j)}{\sum_i^k \sum_j^k \sigma(x_i, x_j)}\right]^2$$

$$= \left\{ \frac{k\sigma_x^2 \left[ 1 + (k-1) \; {}_{12} \right]}{k\sigma_x^2 \left[ 1/ \; {}_{11} + (k-1) \; {}_{12} \right]} \right\}^2$$

$$= \left[ \frac{1 + (k-1) \; {}_{12}}{1/ \; {}_{11} + (k-1) \; {}_{12}} \right]^2 \qquad (7.11)$$

et dans le cas de variables parallèles, elle est égale à

$$\rho^2(D_1^2, d_1^2) = \left[ \frac{k\rho_{11}}{1 + (k-1)\rho_{11}} \right]^2 \qquad (7.12)$$

comme il était normal de s'y attendre, compte tenu de (7.10).

## 3. Les corrélations entre les deux types de distances

Les deux types de distances discutés jusqu'ici donnent lieu à quatre covariances ou corrélations. Selon que les distances sont considérées au niveau vrai ou au niveau observé. Premièrement,

$$\begin{aligned}
\sigma(D_k^2, D_1^2) &= \sigma\left[\Sigma D_i^2, (\Sigma D_i)^2\right] \\
&= \sum_1^k \sigma\left[D_i^2, (\sum_j^k D_j)^2\right] \\
&= \sum_1^k 2\sigma^2\left[D_i, \sum_j^k D\right] \\
&= 2\sum_1^k \left[\sum_j^k \sigma(D_i, D_j)\right]^2 \\
&= 2\sum_1^k \left[\sum_j^i 2\sigma(X_i, X_j)\right]^2 \\
&= 8\sum_1^k \left[\sum_j^k \sigma(X_i, X_j)\right]^2 \qquad (7.13)
\end{aligned}$$

La corrélation suit immédiatement des équations (7.13), (6.1) et (7.5).

$$\rho(D_k^2, D_l^2) = \sigma(D_k^2, D_l^2) \,/\, \sigma(D_k^2)\,\sigma(D_l^2)$$

$$= \frac{\sum_i^k \left[\sum_j^k \sigma(X_i, X_j)\right]^2}{\left[\sum_i^k \sum_j^k \sigma^2(X_i, X_j)\right]^{1/2} \sum_i^k \sum_j^k \sigma(X_i, X_j)} \qquad (7.14)$$

Dans le cas de variables comparables, la covariance précédente est alors égale à

$$\sigma(D_k^2, D_l^2) = 8\sum_i^k \left[\sum_j^k \sigma(X_i, X_j)\right]^2 \qquad (7.13)$$

$$= 8k\left[\sigma_x^2 + (k-1)\sigma_x^2\rho_{12}\right]^2$$

$$= 8k\,\sigma_x^4\left[1 + (k-1)\rho_{12}\right]^2 \qquad (7.15)$$

Mais, puisque, dans le cas de variables comparables,

$$\sigma^2(D_l^2) = 8\left[\sum_{ij}^{kk} \sigma(X_i, X_j)\right]^2$$

$$= 8\left[k\sigma_x^2 + (k-1)k\sigma_x^2\rho_{12}\right]^2$$

$$= 8k^2\sigma_x^4\left[1 + (k-1)\rho_{12}\right]^2 \qquad (7.16)$$

il suit de (7.15) et (7.16) que

$$\sigma(D_k^2, D_l^2) = \sigma^2(D_l^2) \,/\, k \qquad (7.17)$$

La corrélation entre les distances vraies de l'un et l'autre type sera donc égale à

$$\rho(D_k^2, D_l^2) = \frac{\sigma^2(D_l^2)}{k\sigma(D_l^2)\,\sigma(D_k^2)}$$

$$= \sigma(D_l^2) \,/\, k\sigma(D_k^2) \qquad (7.18)$$

$$= \left\{ \frac{8k^2\sigma_x^4 \left[1 + (k-1)\rho_{12}\right]^2}{k_8^2\ k\sigma_x^4 \left[1 + (k-1)\rho_{12}^2\right]} \right\}^{1/2}$$

$$= \frac{1 + (k-1)\rho_{12}}{\left\{k\left[1 + (k-1)\rho_{12}^2\right]\right\}^{1/2}} \qquad (7.19)$$

Dans le cas de variables parallèles, la corrélation devient égale à 1.0, comme il était normal de s'y attendre. Ainsi

$$\rho(D_k^2,\ D_1^2) = \frac{1 + (k-1)\ 1}{\left\{k\left[1 + (k-1)\ 1\right]\right\}^{1/2}} = \frac{k}{(k^2)^{1/2}} = 1.0 \qquad (7.20)$$

Ce résultat s'explique aisément par la corrélation parfaite entre scores vrais dans le cas de variables parallèles. Ainsi $D_k^2 = \sum_1^k D_i^2 = kD_i^2$ et $D_1^2 = (\Sigma D_i)^2 = (kD_i)^2 = k^2 D_i^2 = kD_k^2$.

Deuxièmement, la covariance entre $D_k^2$ et $d_i^2$ emprunte un développement en grande partie semblable à celui de $\sigma(D_k^2,\ D_1^2)$ (cf équation 7.13)

$$\sigma(D_k^2,\ d_1^2) = 2\sum_i^k \left[\sum_j^k \sigma(D_i,\ d_j)\right]^2$$

$$= 2\sum_i^k \left[\sum_j^k \sigma(D_i,\ D_j)\right]^2$$

puisque, suite à la définition 3.3, $\sigma(D_i, d_j) = \sigma(D_i, D_j)$ pour tout i et tout j. Mais cette dernière expression est elle-même égale à $\sigma(D_k^2, D_1^2)$, comme on peut le remarquer à la quatrième ligne du développement de (7.13). Alors

$$\sigma(D_k^2,\ d_1^2) = \sigma(D_k^2,\ D_1^2) \qquad (7.21)$$

La corrélation suit directement des équations (7.21) et (7.9).

$$\rho(D_k^2,\ d_l^2) = \sigma(D_k^2,\ d_l^2) \,/\, \sigma(D_k^2)\ \sigma(d_l^2)$$

$$= \frac{\sigma(D_k^2,\ D_l^2)}{\sigma(D_k^2)\ \sigma(d_l^2)} \cdot \frac{\sigma(D_l^2)}{\sigma(D_l^2)}$$

$$= \rho(D_k^2,\ D_l^2)\,\rho(D_l^2,\ d_l^2) \qquad (7.22)$$

Ce résultat montre que la formule bien connue "d'atténuation" s'applique également aux scores de distances $D_k^2$ et $d_l^2$.

Dans le cas de variables comparables, cette corrélation devient égale, par (7.19) et (7.11), à

$$\rho(D_k^2,\ d_l^2) = \frac{1 + (k-1)\rho_{12}}{\left\{k\left[1 + (k-1)\rho_{12}^2\right]\right\}^{1/2}} \cdot \frac{1 + (k-1)\rho_{12}}{1/\rho_{11} + (k-1)\rho_{12}}$$

$$= \frac{\left[1 + (k-1)\rho_{12}\right]^2}{\left[1/\rho_{11} + (k-1)\rho_{12}\right]\left\{k\left[1 + (k-1)\rho_{12}^2\right]\right\}^{1/2}} \qquad (7.23)$$

Dans le cas de variables parallèles, la corrélation entre $D_k^2$ et $d_l^2$ devient égale à la racine carrée de la fidélité des $d_l^2$, puisque, par (7.20), $\rho(D_k^2,\ D_l^2) = 1.0$

$$\rho(D_k^2,\ d_l^2) = \rho(D_l^2,\ d_l^2) \qquad (7.24)$$

Troisièmement, la covariance entre $d_k^2$ et $D_l^2$ s'obtient par un cheminement identique à celui de l'équation (7.21) et elle est donc égale à

$$\sigma(d_k^2,\ D_l^2) = \sigma(D_k^2,\ D_l^2) \qquad (7.25)$$

La corrélation s'obtient également par un cheminement identique à celui de l'équation (7.22).

$$\rho(d_k^2,\ D_l^2) = \rho(D_k^2,\ D_l^2)\,\rho(D_k^2,\ d_k^2) \tag{7.26}$$

Dans le cas de variables comparables, cette corrélation devient égale, par (7.19 et 7.7), à

$$\rho(d_k^2,\ D_l^2) = \frac{1+(k-1)\rho_{12}}{\left\{k\left[1+(k-1)\rho_{12}^2\right]\right\}^{1/2}}\left[\frac{1+(k-1)\rho_{12}^2}{1/\rho_{11}^2+(k-1)\rho_{12}^2}\right]^{1/2}$$

$$= \frac{1+(k-1)\rho_{12}}{\left\{k\left[1/\rho_{11}^2+(k-1)\rho_{12}^2\right]\right\}^{1/2}} \tag{7.27}$$

Dans le cas de variables parallèles, cette corrélation devient égale à la racine carrée de la fidélité des $d_k^2$, puisqu'il a été démontré, par (7.20), que la corrélation entre les distances vraies de l'un et l'autre type est ici parfaite.

$$\rho(d_k^2,\ D_l^2) = \rho(d_k^2,\ D_k^2) \tag{7.28}$$

Quatrièmement et dernièrement, la covariance entre distances observées de l'un et l'autre type s'obtient de façon similaire à (7.13), où il suffit de remplacer les variables vraies $D_i$ et $X_i$ par les variables observées $d_i$ et $x_i$. Elle est donc égale à

$$\sigma(d_k^2,\ d_l^2) = 8\sum_{i}^{k}\left[\sum_{j}^{k}\sigma(x_i,\ x_j)\right]^2 \tag{7.29}$$

$$= 8\sum_{i}^{k}\left[\sigma^2(x_i)+\sum_{i\neq j}^{k}\sigma(x_i,\ x_j)\right]^2$$

$$= 8\sum_{i}^{k}\left[\sigma^2(X_i)+\sigma^2(E_i)+\sum_{i\neq j}^{k}\sigma(X_i,\ X_j)\right]^2$$

$$= 8\sum_{i}^{k}\left[\sum_{j}^{k}\sigma(X_i,\ X_j)+\sigma^2(E_i)\right]^2 \tag{7.30}$$

La comparaison de (7.30) à (7.13) indique que la covariance entre ces distances observées, $\sigma(d_k^2,\ d_l^2)$ sera toujours supérieure à la covariance entre les distances vraies correspondantes, $\sigma(D_k^2,\ D_l^2)$.

Ce résultat inhabituel s'explique aisément par le fait que les erreurs de mesures présentes dans $d_k^2$ se retrouvent également dans $d_1^2$. Quant à la corrélation, elle suit directement des équations (7.29), (7.4) et (7.6).

$$\rho(d_k^2, d_1^2) = \sigma(d_k^2, d_1^2) \,/\, \sigma(d_k^2)\,\sigma(d_1^2)$$

$$= \frac{\sum_i^k \left[\sum_j^k \sigma(x_i, x_j)\right]^2}{\left[\sum_{ij}^{kk} \sigma^2(x_i, x_j)\right]^{1/2} \sum_{ij}^{kk} \sigma(x_i, x_j)} \qquad (7.31)$$

Dans le cas de variables comparables, la covariance de (7.29) devient égale à

$$\sigma(d_k^2, d_1^2) = 8\sum_i^k \left[\sigma^2(x_i) + \sum_{i\neq j}^k \sigma(x_i, x_j)\right]^2$$

$$= 8k\sigma_x^4 \left[1/ \quad_{11} + (k-1) \quad_{12}\right]^2$$

$$= \sigma^2(d_1^2) \,/\, k \qquad (7.32)$$

puisque, dans le cas de variables comparables, la variance des $d_1^2$ (équation 7.6) est elle-même égale à

$$\sigma^2(d_1^2) = 8\left[\sum_{ij}^{kk} \sigma(x_i, x_j)\right]^2 \qquad (7.6)$$

$$= 8k^2\sigma_x^4 \left[1/\rho_{11} + (k-1)\rho_{12}\right]^2 \qquad (7.33)$$

Mais compte tenu que, dans ce cas particulier, la variance des $d_k^2$ est égale à

$$\sigma^2(d_k^2) = 8\sum\sum\sigma^2(x_i, x_j) \qquad (7.4)$$

$$= 8k\sigma_x^4 \left[1/\rho_{11}^2 + (k-1)\rho_{12}^2\right] \qquad (7.34)$$

la corrélation sera donc égale, par (7.32) à

$$\rho(d_k^2, d_1^2) = \sigma(d_k^2, d_1^2) \,/\, \sigma(d_k^2)\,\sigma(d_1^2)$$

$$= \sigma^2(d_1^2) \,/\, k\sigma(d_k^2)\,\sigma(d_1^2)$$

$$= \sigma(d_1^2) \,/\, k\sigma(d_k^2) \qquad (7.35)$$

et par (7.33 et (7.34),

$$= \frac{1/_{11} + (k-1)\rho_{12}}{\left\{k\left[1/\rho_{11}^2 + (k-1)\rho_{12}^2\right]\right\}^{1/2}} \qquad (7.36$$

Dans le cas de variables parallèles, cette corrélation se simplifie en

$$\rho(d_k^2, d_1^2) = \frac{1/\rho_{11} + (k-1)}{\left\{k\left[1/\rho_{11}^2 + (k-1)\right]\right\}^{1/2}}$$

$$= \frac{1 + (k-1)\rho_{11}}{\left\{k\left[1 + (k-1)\rho_{11}^2\right]\right\}^{1/2}} \qquad (7.37)$$

Tableau 7.1

Résumé des formules des études 6 et 7

| Statistiques | Variables quelconques | | Variables comparables | | Variables parallèles | |
|---|---|---|---|---|---|---|
| $\sigma^2(D_K^2)$ | $8\sum_i^n\sum_j^n \sigma^2(X_i, X_j)$ | (6.1) | $8K\sigma_X^4\left[1+(K-1)\rho_{12}^2\right]$ | | $8K^2\sigma_X^4$ | |
| $\sigma^2(d_K^2)$ | $8\sum_i^n\sum_j^n \sigma^2(x_i, x_j)$ | (7.4) | $8K\sigma_X^4\left[1/\rho_{11}^2+(K-1)\rho_{12}^2\right]$ | (7.34) | | |
| $\sigma(D_K^2, d_K^2)$ | $\sigma^2(D_K^2)$ | (6.11) | | | | |
| $\rho^2(D_K^2, d_K^2)$ | $\dfrac{\sigma^2(D_K^2)}{\sigma^2(d_K^2)}$ | (6.12) | $\dfrac{1+(K-1)\rho_{12}^2}{1/\rho_{11}^2+(K-1)\rho_{12}^2}$ | (7.7) | $\dfrac{K\rho_{11}^2}{1+(K-1)\rho_{11}^2}$ | (6.15) |
| $\sigma^2(D_1^2)$ | $8\left[\sum_i^n\sum_j^n \sigma(X_i, X_j)\right]^2$ | (7.5) | $8K^2\sigma_X^4\left[1+(K-1)\rho_{12}\right]^2$ | | $8K^4\sigma_X^4$ | |
| $\sigma^2(d_1^2)$ | $8\left[\sum_i^n\sum_j^n \sigma(x_i, x_j)\right]^2$ | (7.6) | $8K^2\sigma_X^4\left[1/\rho_{11}+(K-1)\rho_{12}\right]^2$ | (7.33) | | |
| $\sigma(D_1^2, d_1^2)$ | $\sigma^2(D_1^2)$ | (7.8) | | | | |
| $\rho^2(D_1^2, d_1^2)$ | $\dfrac{\sigma^2(D_1^2)}{\sigma^2(d_1^2)}$ | (7.9) | $\left[\dfrac{1+(K-1)\rho_{12}}{1/\rho_{11}+(K-1)\rho_{12}}\right]^2$ | (7.11) | $\left[\dfrac{K\rho_{11}}{1+(K-1)\rho_{11}}\right]^2$ | (7.12) |

| | | | | | | |
|---|---|---|---|---|---|---|
| $\sigma(D_K^2, D_i^2)$ | $8\sum_i^K\left[\sum_j^K \sigma(X_i, X_j)\right]^2$ | (7.13) | $\sigma^2(D_i^2)/K$ | (7.17) | $8K^3\sigma_x^4$ | |
| $\rho(D_K^2, D_i^2)$ | $\dfrac{\sigma(D_K^2, D_i^2)}{\sigma(D_K^2)\,\sigma(D_i^2)}$ | (7.14) | $\dfrac{1+(K-1)\rho_{12}}{\left\{K\left[1+(K-1)\rho_{12}^2\right]\right\}^{1/2}}$ | (7.19) | 1.0 | (7.20) |
| $\sigma(D_K^2, d_i^2)$ | $\sigma(D_K^2, D_i^2)$ | (7.21) | | | | |
| $\rho(D_K^2, d_i^2)$ | $\rho(D_K^2, D_i^2)\rho(D_i^2, d_i^2)$ | (7.22) | $\dfrac{\left[1+(K-1)\rho_{12}\right]^2}{\left\{K\left[1+(K-1)\rho_{12}^2\right]\right\}^{1/2}\left[1/\rho_{11}+(K-1)\rho_{12}\right]}$ | (7.23) | $\rho(D_i^2, d_i^2)$ | (7.24) |
| $\sigma(d_K^2, D_i^2)$ | $\sigma(D_K^2, D_i^2)$ | (7.25) | | | | |
| $\rho(d_K^2, D_i^2)$ | $\rho(D_K^2, D_i^2)\rho(D_K^2, d_K^2)$ | (7.26) | $\dfrac{1+(K-1)\rho_{12}}{\left\{K\left[1/\rho_{11}^2+(K-1)\rho_{12}^2\right]\right\}^{1/2}}$ | (7.27) | $\rho(D_K^2, d_K^2)$ | (7.28) |
| $\sigma(d_K^2, d_i^2)$ | $8\sum_i^K\left[\sum_j^K \sigma(\alpha_i, \alpha_j)\right]^2$ | (7.29) | $\sigma^2(d_i^2)/K$ | (7.32) | | |
| $\rho(d_K^2, d_i^2)$ | $\dfrac{\sigma(d_i^2)}{K\,\sigma(d_K^2)}$ | (7.35) | $\dfrac{1/\rho_{11}+(K-1)\rho_{12}}{\left\{K\left[1/\rho_{11}^2+(K-1)\rho_{12}^2\right]\right\}^{1/2}}$ | (7.36) | $\dfrac{1+(K-1)\rho_{11}}{\left\{K\left[1+(K-1)\rho_{11}^2\right]\right\}^{1/2}}$ | (7.37) |

Appendice A

```
PROGRAM  ETUDES                                        CDC 6400 FTN V3.0-P365 OP
      PROGRAM ETUDES(OUTPUT)
      DIMENSION    Z(50,5),X(50,2),XO(50,2),XT(50,1),
     C XO(50,50),YD(50,50),R(5),RR(2,5,50),XM(2,5),XSC(2,5)
      C=2
    5 R(1)=0 $ R(2)=.25$ R(3)=.5 $ R(4)=.75 $ R(5)=1
      DO 5 L=1,5
      A=SQRT(R(L)) $ B=SQRT(1-R(L))
      CALL RANSET(66)
      DO 5 M=1,50
   10 DO1I=1,50
      DO 1 J=1,5
      Z(I,J)=-6
      DO 1 K=1,12
    1 Z(I,J)=Z(I,J)+RANF(0)
   15 DO 2 I=1,50
      X(I,1)=A*Z(I,1)+B*Z(I,2)
      X(I,2)=A*Z(I,1)+B*Z(I,3)
      XO(I,1)=X(I,1)+C*Z(I,4)
      XO(I,2)=X(I,2)+C*Z(I,5)
   20 XT(I,1)=XO(I,1)+XO(I,2)
    2 CONTINUE
      IF(L.GT.1. OR.M.GT.1)GOTO 15
      PRINT 7
      DO 14 I=1,50
   25 14 PRINT 13,I,X(I,1),X(I,2),XO(I,1),XO(I,2),XT(I,1)
   13 FORMAT(I3,2(2X,2F6.2),2X,F6.2)
   15 CONTINUE
      CALL MATDIST(X,50,2,XO)
      CALL MATDIST(XO,50,2,YD)
   30 CALL CORDIST(XO,YD,50,RR(1,L,M))
      CALL MATDIST(X1,50,1,YD)
      CALL CORDIST(XO,YD,50,RR(2,L,M))
    5 CONTINUE
      PRINT 7
   35 7 FORMAT(1H1///)
      DO 4 M=1,50
    4 PRINT 6,M,((RR(K,L,M),K=1,2),L=1,5)
    6 FORMAT(I3,5(3X,2F7.3))
      DO 8 K=1,2
   40 DO 8 L=1,5
      DO 8 M=1,50
    8 RR(K,L,M)=.5*ALOG((1+RR(K,L,M))/(1-RR(K,L,M)))
      PRINT 7
      DO 11 M=1,50
   45 11 PRINT 6,M,((RR(K,L,M),K=1,2),L=1,5)
      DO 9 K=1,2
      DO 9 L=1,5
      S=SC=0
      DO 10 M=1,50
   50 S=S+RR(K,L,M)
   10 SC=SC+RR(K,L,M)**2
      XM(K,L)=S
    9 XSC(K,L)=SC-(S**2)/50
      PRINT 6,M,((XM(K,L),K=1,2),L=1,5)
   55 PRINT 6,M,((XSC(K,L),K=1,2),L=1,5)
```

```
PROGRAM    WARRING                                    CDC 6400 FTN V3.0-P365 OPT

      PROGRAM WARRING(OUTPUT)
      DIMENSION Z(600),T(50,2),X(50,10),XS(50,2),XT(50,1),
     C XD(50,50),YD(50,50)
      B=1
 5    CALL RANSET (74)
      DO 10 NFOIS=1,20
      DO 1 I=1,600
      Z(I)=-6.0
      DO 1 J=1,12
10    ZA=RANF(0)
    1 Z(I)=Z(I)+ZA
      DO 2 I=1,50
      J=I*12-11
      T(I,1)=Z(J) $ T(I,2)=Z(J+1)
15    XS(I,1)=0
      DO 3 K=1,5
      X(I,K)=T(I,1)+B*Z(J+1+K)
    3 XS(I,1)=XS(I,1)+X(I,K)
      XS(I,2)=0
20    DO 4 K=6,10
      X(I,K)=T(I,2)+B*Z(J+1+K)
    4 XS(I,2)=XS(I,2)+X(I,K)
    2 XT(I,1)=XS(I,1)+XS(I,2)
      IF(NFOIS.GT.1) GOTO 9
25    PRINT 7
    7 FORMAT(1H1)
      DO 6 I=1,50
    6 PRINT 5, I,T(I,1),T(I,2),(X(I,J),J=1,10),XS(I,1),XS(I,2)
    5 FORMAT(I3,5X,2F8.3,5X,10F8.3,5X,2F8.3)
30    9 CONTINUE
      CALL MATDIST(T,50,2,XD)
      CALL MATDIST(X,50,10,YD)
      CALL CORDIST(XD,YD,50,R)
      CALL MATDIST(XS,50,2,YD)
35    CALL CORDIST(XD,YD,50,R)
      CALL MATDIST(XT,50,1,YD)
      CALL CORDIST(XD,YD,50,R)
   10 CONTINUE
      STOP $ END
```

```
      SUBROUTINE MATDIST(X,N,NV,XD)
      DIMENSION X(50,10),XD(50,50)
      NM1=N-1
      DO 1 I=1,NM1
    5 IP1=I+1
      DO 1 J=IP1,N
      XD(I,J)=0
      DO 2 K=1,NV
    2 XD(I,J)=XD(I,J)+(X(I,K)-X(J,K))**2
   10 1 XD(I,J)=SQRT(XD(I,J))
      RETURN $ END
```

```
      SUBROUTINE CORDIST(XD,YD,N,R)
      DIMENSION XD(50,50),YD(50,50)
      M=N*(N-1)/2
      SX=SY=SXX=SYY=SXY=0
      NM1=N-1
      DO 1 I=1,NM1
      IP1=I+1
      DO 1 J=IP1,N
      SX=SX+XD(I,J) $ SY=SY+YD(I,J)
      SYY=SYY+YD(I,J)**2 $SXX=SXX+XD(I,J)**2
    1 SXY=SXY+XD(I,J)*YD(I,J)
      R=(M*SXY-SX*SY)/SQRT((M*SXX-SX**2)*(M*SYY-SY**2))
      PRINT 2,R
    2 FORMAT(10X,*R=*,F10.7)
      RETURN $ END
```

Appendice B

```
PROGRAM      ETUDES                                    CDC 6400 FTN V3.0-P365 OP
                     DO 16 K=1,2
                     DO 16 L=1,5
                     E=EXP(XM(K,L)/25)
                  16 XM(K,L)=(E-1)/(E+1)
      60             PRINT 6,M,((XM( K,L),K=1,2),L=1,5)
                     STOP $ END
```

**Notes:** 1. Ce programme a été exécuté 6 fois.
L'instruction de la quatrième ligne a été modifiée trois fois
(C = .5; C = 1; C = 2)
L'instruction de la cinquième ligne a été modifiée deux fois
(R = .00, .25, .50, .75, 1.00; R = .125, .375, .625, .875, 1.00)

2. Les sous-routines MATDIST et CORDIST ont déjà présentées à l'Appendice A.

3. Le temps total d'exécution a été de 825 secondes.

## Références


CARROLL, J.B. (1961) The nature of data, and how to choose a correlation coefficient. Psychometrika, 26, 347-372.

CARROLL, R.M. et FIELD, J. (1974) A comparison of the classification accuracy of profile similarity measures. Multivariate Behavioral Research, , 373-380.

CRONBACH, L.J. (1953) Correlations between persons as a research tool. In O.H. Mowrer (Ed.) Psychotherapy: Theory and research. New York: Ronald Press.

CRONBACH, L.J. et GLESER, G.L. (1953) Assessing similarity between profiles. Psychological Bulletin, 50(6), 456-473.

HELMSTADTER, G.C. (1957) An empirical comparison of methods for estimating profile similarity. Educational and psychological Measurement, 17, 71-82.

LORD, F.M. et NOVICK, M.R. (1968) Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

MOSEL, J.N. et ROBERTS, J.B. (1954) The comparability of measures of profile similarity: an empirical study. Journal of Consulting Psychology, 18, 61-66.

Université du Québec
**Institut national de la recherche scientifique**